ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

CREDIT SCORING: APLICANDO TÉCNICAS DE REGRESIÓN LOGÍSTICA Y MODELOS ADITIVOS GENERALIZADOS PARA UNA CARTERA DE CRÉDITO EN UNA ENTIDAD FINANCIERA.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

JAIME ANDRÉS SUQUILLO LLUMIQUINGA

jaimhe_13mh@hotmail.com

DIRECTORA: ADRIANA UQUILLAS ANDRADE, PhD.

adriana.uquillas@epn.edu.ec

Quito, Octubre de 2021

DECLARACIÓN

Yo, Jaime Andrés Suquillo Llumiquinga, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

JAIME ANDRÉS SUQUILLO LLUMIQUINGA

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Jaime Andrés Suquillo LLumiquinga, bajo mi supervisión.

_____

Adriana Uquillas, PhD.
DIRECTOR DE PROYECTO

AGRADECIMIENTOS

La vida está llena de matices y de destellos de felicidad. En todo este proceso he vivido, crecido, aprendido y mejorado.

Soy consciente de que esto no hubiese sido posible sin el apoyo de mi amada madre Blanquita, sus consejos y ejemplo han hecho de mi un hombre con fuertes principios y valiosas virtudes. Gracias a mis hermanas que nunca me dejaron solo.

A mi padre Jaime en el cielo que, aunque no lo haya conocido lo llevo en mi corazón, es mi fuerza y mi inspiración.

A mis abuelitos Lucrecia, Manuel, Alfredo y Mercedes; a mi tío Bolívar y padrinos Flavio y Nelly que amo y que han sido mis segundos padres.

A Diana, la mujer con la que he crecido y vivido todo este proceso. Gracias por tu amor y apoyo.

Gracias a mi Directora de Tesis Dra. Adriana Uquillas por confiar en mí y por brindarme la oportunidad de tener un crecimiento tanto profesional como intelectual.

A todas las personas que han formado parte de este gran proceso, profesores, amigos y familia.

DEDICATORIA

A la memoria de mi padre Jaime. Él es un león que me cuida desde el cielo en cada paso que doy.

A mi amada madre Blanquita, a quien van dedicados todos mis logros. Ella lo merece todo.

Índice general

Índice de figuras

Índice de tablas

Resumen

El presente proyecto de titulación tiene como finalidad el estudio de una metodología estadística basada en medidas de asociación, medidas de separación y presentar el Modelo Aditivo Generalizado como una alternativa prometedora a la Regresión Logística, los cuales se aplican en la construcción de un modelo estadístico, que permita estimar la probabilidad de incumplimiento (PD) de una cartera de algún tipo de crédito consedido.

La regresión logística es el modelo estadístico más utilizado en la industria de calificación crediticia. A pesar de sus ventajas en la fácil interpretación y el bajo costo computacional, la regresión logística está bajo la crítica de la dificultad de modelar las características no lineales del efecto de los predictores sobre la variable dependiente y, por lo tanto, podría dar lugar a resultados insatisfactorios.

En el presente estudio se plantea la utilización de una técnica conocida como Modelo Aditivo Generalizado introducido por Hastie y Tibshirani (1990), que proporciona la capacidad de detectar la relación no lineal y no monotónica entre la variable dependiente y los predictores sin sacrificar la interpretabilidad.

El rendimiento de los modelos se evalúan y comparan utilizando el valor del estadístico de Kolmogórov - Smirnov (KS), el área bajo la curva de Características operativas del receptor (ROC) y el test de GINI.

Adicionalmente, haciendo uso de software estadístico R, se implementan las técnicas descritas, el cual permite obtener los resultados de manera automática disminuyendo considerablemente el tiempo empleado.

Palabras claves: Medidas de separación, medidas de asociación, regresión logística, modelo logístico aditivo generalizado, programación en R.

Abstract

The purpose of this degree project is to study a statistical methodology based on association measures, separation measures and present the Generalized Additive Model as a promising alternative to Logistic Regression, which are applied in the construction of a statistical model, that allows estimating the probability of default (PD) of a portfolio of some type of granted credit.

Logistic regression is the most used statistical model in the credit rating industry. Despite its advantages in easy interpretation and low computational cost, logistic regression is under the criticism of the difficulty of modeling the non-linear characteristics of the predictors' effect on the dependent variable and, therefore, could lead to Unsatisfactory results.

The present study proposes the use of a technique known as Generalized Additive Model introduced by Hastie and Tibshirani (1990), which provides the ability to detect the non-linear and non-monotonic relationship between the dependent variable and the non-sacrificing predictors Interpretability

The performance of the models are evaluated and compared using the Kolmogórov - Smirnov (KS) statistic value, the area under the Receiver Operating Characteristics (ROC) curve and the GINI test.

Additionally, making use of statistical software R, the techniques described are implemented, which allows obtaining the results automatically, considerably reducing the time used.

Key words: Separation measures, association measures, logistic regression, generalized additive logistic model, R programming.

Capítulo 1

Introducción.

La concesión de crédito en la actualidad es una de las principales operaciones realizadas en diversas áreas tanto financieras como comerciales.

[Anderson, 2007] señala que:

> En el contexto actual, “crédito” simplemente significa “compre ahora, pague después”, ya sea que la compra sea para consumo a corto plazo, bienes duraderos y otros activos que brinden a los usuarios servicios valiosos o empresas productivas. La palabra “crédito” proviene de la antigua palabra latina “credo”, que significa “confianza en” o “confiar en”. Si le prestas algo a alguien, debes confiar en él para cumplir con la obligación (p.62)

Según el Diccionario de la Real Academia de la Lengua Española, crédito se define como:

- m. Cantidad de dinero u otro medio de pago que una persona o entidad, especialmente bancaria presta a otro bajo determinadas condiciones de devolución.
- m. Apoyo, abono, comprobación.
- m. Situación económica o condiciones morales que facultan a una persona o entidad para obtener de otra fondos o mercancías.

Por el contrario, el Diccionario Jurídico, define a crédito como:

- Contrato por el cual una persona natural o jurídica obtiene temporalmente una cantidad de dinero de otra a cambio de una remuneración en forma de intereses. Llegado

el momento del vencimiento, el deudor deberá devolver el monto otorgado más sus respectivos intereses.

La concesión de créditos como es de esperar representa diferentes tipos de riesgos destacando el riesgo de crédito que pueden poner en peligro la estabilidad de las instituciones financieras afectando su liquidez y solvencia.

Según la [Superintendencia de Bancos y Seguros, Libro 1]:

2.1 Riesgo.- Es la posibilidad de que se produzca un hecho generador de pérdidas que afecten el valor económico de las instituciones;

[...]

2.4 Riesgo de crédito.- Es la posibilidad de pérdida debido al incumplimiento del prestatario o la contraparte en operaciones directas, indirectas o de derivados que conlleva el no pago, el pago parcial o la falta de oportunidad en el pago de las obligaciones pactadas;

[Aguilar y Camargo, 2004] al respecto señalan: “un elevado número de créditos en condición de retraso o de no pago constituyen una de las principales causas de la insolvencia y descapitalización; lo que finalmente atenta contra la solidez y sostenimiento de la institución en el largo plazo”.

En efecto, la fragilidad de una institución financiera debido a altos niveles de morosidad de sus créditos conlleva inicialmente a un problema de liquidez, que, en el largo plazo, si es recurrente y si la institución no posee líneas de créditos de contingencia, se convierte en uno de solvencia que, que determina, probablemente, la liquidación de la institución (Freixas y Rochet, 1998 cit. en [Aguilar y Camargo, 2004]).

Podemos definir el sistema financiero como un conjunto de instituciones, instrumentos y mercados a través de los cuales se canaliza el ahorro hacia la inversión. Este ahorro será canalizado desde las unidades excedentarias (prestamistas) hacia las unidades deficitarias (prestatarios), mediante la intervención de una serie de intermediarios financieros [González y López, 2008].

Esto sin duda aporta significativamente al crecimiento económico de un país, dependiendo de las particularidades de cada una de las economías. Siguiendo esta línea, el buen funcionamiento, así como la sostenibilidad de las instituciones financieras constituyen un papel fundamental dentro del sistema financiero. Una de las actividades principales que ayudan a cumplir con este fin, es el adecuado manejo del riesgo de crédito.

La [Superintendencia de Bancos y Seguros, Libro 1], señala que: “Cada entidad controlada tiene su perfil de riesgo según las características de los mercados en los que opera y de los productos que ofrece; por tanto, al no existir un modelo único de administración del riesgo de crédito cada entidad debe desarrollar su propio esquema.”

Por tal motivo, las instituciones financieras, con el objetivo de reducir el riesgo de incumplimiento asociados a la concesión de crédito se han visto en la necesidad de implementar diferentes mecanismos que permitan evaluar la calificación crediticia de los clientes, distinguiendo a los “buenos” de los “malos”, pronosticando y previniendo pérdidas futuras frente al incumplimiento de pago por parte del cliente.

El credit scoring es un sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. Es capaz de predecir la probabilidad de no pago de los clientes con sus obligaciones y la severidad de las pérdidas en caso de incumplimiento, asociada a una operación crediticia, estos son componentes claves para determinar el riesgo de crédito de una cartera [Elizondo, 2003].

Los modelos estadísticos, como el análisis discriminante lineal, la regresión logística, el árbol de clasificación y regresión y la red neuronal, se utilizan ampliamente para evaluar la solvencia crediticia de los prestatarios potenciales a fin de reducir el riesgo de incumplimiento (Franke, Hardle y Stahl 2000, Shao 2004).

La regresión logística, que es un caso especial de los modelos lineales generalizados, es el modelo estadístico más utilizado en la industria de calificación crediticia. A pesar de sus ventajas en la fácil interpretación y el bajo costo computacional, la regresión

logística está bajo la crítica de la dificultad de modelar las características no lineales del efecto de los predictores sobre la variable dependiente y, por lo tanto, podría dar lugar a resultados insatisfactorios. [Liu and Cela, 2007] dicen: “las técnicas estadísticas modernas, como la red neuronal y la regresión de búsqueda de proyección, han demostrado ser exitosas en el modelado no lineal. Sin embargo, este éxito viene con el precio de la interpretabilidad”.

En el presente estudio se plantea la utilización de una técnica conocida como Modelo Aditivo Generalizado introducido por [Hastie and Tibshirani, 1990] y que proporciona la capacidad de detectar los patrones no lineales sin sacrificar la interpretabilidad. Para este estudio se utilizará un ejemplo específico de los modelos aditivos generalizados: una generalización del modelo logístico *(logit)* para valores de variables dependientes binarias conocido como Modelo Logístico Aditivo Generalizado.

La información necesaria para el desarrollo de la metodología estadística, para fines académicos, será proporcionada por una institución financiera de un país emergente. Se dispondrá principalmente de información histórica relacionada con aspectos sociodemográficos, hábito de pago y consumo del cliente en entidades del sistema financiero.

Este estudio busca hacer notar la existencia de una técnica estadística moderna y confiable. De este modo podremos comparar empíricamente los resultados obtenidos mediante la técnica regresión logística y el modelo logístico aditivo generalizado con la finalidad de mostrar su mayor precisión e interpretación.

Con el fin de facilitar la utilización y aplicación en cualquier entidad financiera, es necesario disponer de un algoritmo informático que permita ejecutar automáticamente la metodología desarrollada. El algoritmo que generemos recibirá como entrada nuestra base de datos con la información histórica y una marca binaria que identifica $(Bueno/Malo)$, presentará el modelo estadístico ajustado y los resultados necesarios para validarlo.

1.1 Modelos estadísticos usados en Credit Scoring

En la actualidad, para evaluar el riesgo de crédito o a su vez la conveniencia en la concesión de un crédito, existe gran variedad de metodologías con diferentes enfoques (ver Srinivasan y Kim (1987), Mester(1997), Hand y Henley (1997) y Thomas (2000)): análisis discriminante, regresión lineal, regresión logística, modelos probit, modelos logit, métodos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, algoritmos de particionamiento recursivo (árboles de decisión), sistemas expertos, algoritmos genéticos y redes neuronales [Girault, 2007].

La Regresión Logística ocupa una posición central en el campo de la calificación crediticia [Kleinbaum, 1994], y en Ecuador es la técnica comúnmente utilizada ya que se entiende relativamente bien y se puede derivar una fórmula explícita en la que se puedan basar las decisiones de crédito. Además, es ampliamente utilizada en la industria y se ha convertido en el estándar empleado por la mayoría de las empresas. Aunque las redes neuronales artificiales pueden ser más poderosas que la regresión logística, no se usan ampliamente en la calificación crediticia porque es una caja negra con respecto a la interpretación y la ausencia de razones por las cuales la red neuronal ha tomado sus decisiones, puede ser inaceptable. A diferencia de la regresión logística y las redes neuronales, “los modelos aditivos generalizados ofrecen un punto medio: pueden adaptarse a relaciones complejas y no lineales y hacer buenas predicciones en estos casos, pero aún podemos hacer estadísticas inferenciales, comprender y explicar la estructura subyacente de nuestros modelos y por qué hacen las predicciones que hacen” [Ross, 2019].

Por lo que tratar de innovar en técnicas que generen mejores y correctos resultados es motivo suficiente para estudiarlo y con ello en un futuro poder implementarlo en la industria pues sería beneficioso para el país, es así que a través de la investigación se ha logrado descubrir métodos más dinámicos, con mayor precisión e interpretación que a su vez generarán mayor rentabilidad, con dichos métodos se ha logrado mejores resultados.

Estudios como los de [Liu and Chuck, 2009] y [Lohmann and Ohliger, 2018] demuestran que la técnica conocida como modelo aditivo generalizado funciona perfectamente en problemas de credit scoring y es una técnica eficaz, sencilla y superior a técnicas paramétricas tradicionalmente utilizadas, entre ellas, la regresión logística.

En este trabajo se propone el uso de modelos aditivos generalizados (GAM), técnica no utilizada en Ecuador; y se realizará una comparación con la regresión logística con la finalidad de mostrar mayor poder predictivo e interpretativo. La regresión logística forma parte de los GLM (Modelos Lineales Generalizados), mientras que los GAM son una generalización de los GLM.

El método estadístico flexible que puede usarse para identificar y caracterizar los efectos de regresión no lineal. Se denomina Modelos Aditivos Generalizados (GAM).

1.1.1 Modelo Aditivo Generalizado

Se presenta una breve introducción del concepto y funcionamiento del modelo, pues en el siguiente capítulo se lo revisará a profundidad. Este modelo es una generalización de los Modelos Lineales Generalizados los que se diferencian porque el predictor lineal ya no es simplemente una combinación lineal de las variables explicativas, sino que es una combinación lineal de funciones de dichas variables explicativas, lo que permite introducir en el modelo todo tipo de efectos y relaciones no lineales entre variables bajo ciertas condiciones proporciona un mejor resultado que el que nos brindaría sin ajustes no lineales, que nos permite el modelo GAM.

El modelo está construido por la suma de funciones suaves (“no paramétricas”) no especificadas de las variables independientes x_i, pudiendo ser estas variables continuas, variables categóricas, número de casos y series de datos. A diferencia de los modelos de regresión lineal donde se deben determinar los parámetros correspondientes a cada uno de los predictores x_i, el modelo sustituye $\sum \beta_i x_i$ por una suma de funciones no determinadas lineales $\sum \beta_i f_i(x_i)$, donde cada una de las f_i es estimada de manera muy flexible, pudiendo mostrar este efecto no lineal de esa relación.

1.2 Descripción de la Metodología Analítica

La información que se utilizará en este estudio son datos relacionados con el comportamiento crediticio histórico, aspectos sociodemográficos y hábitos de pago y consumo de clientes proporcionados por una institución financiera de un país emergente. La información es proporcionada únicamente con fines académicos.

Con la información proporcionada se procede a realizar los siguientes pasos:

- **Tratamiento de los datos:** El rendimiento del modelo que se va a construir dependerá de la calidad de la información que se disponga. Por lo tanto, como primer paso se realiza un análisis descriptivo: calculando medidas de tendencia central y posición para variables independientes cuantitativas, tablas de frecuencia para las variables cualitativas. Además, se realiza un análisis de exactitud y completitud a los datos. Se realiza un análisis y tratamiento de datos atípicos. Por último, se recategorizan tanto las variables cualitativas como cuantitativas.
- **Construcción del modelo:** Para la selección de variables que formarán parte del modelo estadístico, se hará uso de los estadísticos Kolmogórov - Smirnov (KS) y valor de información (VI). Además, se hace uso del método de ajuste o suavizado (splines) a aplicar en las variables continuas que ingresen al modelo logístico aditivo generalizado.
- **Evaluación estadística del modelo:** Para evaluar el rendimiento del modelo estadístico se van a realizar tablas de desempeño y pruebas que permitirán medir características de ajuste y discriminación.

El modelo estadístico óptimo permite obtener resultados que proporcionan herramientas necesarias a una institución financiera que le faciliten en:

- La identificación de características que más influyen en la persona para que se le otorgue el crédito.

- La segmentación de la cartera de clientes de una institución financiera de acuerdo a la probabilidad de incumplimiento (PD).
- La focalización de estrategias en cada segmento minimizando así el riesgo inherente a la probabilidad de incumplimiento.
- El diseño de políticas y procedimientos para mejorar la gestión en la concesión del crédito.

Una vez que se ha planteado el problema actual que involucra al algoritmo de regresión logística sobre la dificultad de modelar las características no lineales del efecto de los predictores sobre la variable dependiente y, por lo tanto, podría dar lugar a resultados insatisfactorios.

En la práctica ocurre que para capturar la no linealidad en los modelos de regresión mediante alguna función f, el experto transforma o categoriza alguno o todos los predictores y así poder modelar fenómenos más complejos. Sin embargo, al transformar matemáticamente una variable se puede perder la relación real existente entre la variable respuesta y las variables explicativas, ya que pueden existir relaciones que tengan una forma desconocida, el modelo logístico aditivo generalizado permite modelar de manera flexible las relaciones no lineales sin realizar ninguna suposición sobre la forma funcional de f, esto ofrece una mejor predicción sin perder su capacidad interpretativa. Por lo tanto, el conjunto de variables explicativas final podría ser diferente si se hace uso de uno u otro modelo. En la práctica el uso de un modelo GAM puede resultar más sencillo.

Desarrollada la metodología analítica, esta se la implementará en el lenguaje de programación estadístico **R**(R Core Team, 2016), con el fin de optimizar su tiempo de ejecución. El algoritmo recibirá como entrada un número determinado de datos con la información histórica y una marca binaria que identifica $(Bueno/Malo)$ y retornará el modelo estadístico ajustado.
R es un lenguaje y entorno para computación estadística y gráficos. Es un proyecto GNU que es similar al lenguaje S y al entorno que fue desarrollado en los Laboratorios Bell $($anteriormente AT&T, ahora Lucent Technologies$)$ por John Chambers y sus

colegas.

R proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, ...) y gráficas, y es altamente extensible.

R es un conjunto integrado de instalaciones de software para la manipulación de datos, el cálculo y la visualización gráfica. Incluye

- una instalación eficaz de manejo y almacenamiento de datos, un conjunto de operadores para cálculos en matrices, en particular matrices,
- una colección grande, coherente e integrada de herramientas intermedias para el análisis de datos,
- facilidades gráficas para el análisis y visualización de datos en pantalla o en papel, y
- un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario e instalaciones de entrada y salida.

Descrita la metodología analítica, el software estadístico y la información disponible. Se procede a fijar los objetivos a cumplir en el presente estudio, cuyo fin es proporcionar una alternativa prometedora para la calificación crediticia.

1.3 Objetivos

1.3.1 Objetivo General

Comparar dos técnicas estadísticas: Regresión Logística y Modelos Aditivos Generalizados que permitirán estimar la probabilidad de incumplimiento de un cliente al momento de la concesión del crédito.

1.3.2 Objetivos Específicos

- Establecer los determinantes del incumplimiento del crédito.
- Establecer modelos estadísticos robustos mediante la técnica Regresión Logística y el algoritmo de Modelos Aditivos Generalizados que permitan comparar empíricamente los resultados obtenidos.
- Implementar un código en el lenguaje de programación estadístico **R** que realice automáticamente la metodología analítica utilizada en la generación del modelo y nos propine resultados óptimos para la toma de decisiones acertadas.

Capítulo 2

Marco Teórico

En el presente capítulo presentaremos los conceptos teóricos necesarios para comprender las metodologías de regresión logística y modelos aditivos generalizados empleadas en la construcción y validación de cada modelo estadístico; y conocer su estructura. Además, se describen estadísticos e índices empleados como criterio de selección de variables que permiten mejorar el rendimiento de discriminación del modelo: medidas de separación y asociación para variables cuantitativas y cualitativas respectivamente.

Una institución financiera cumple como intermediaria financiera, canalizando el ahorro (exceso) de recursos hacia quienes lo necesiten, sean usados estos, para consumo o inversión. Con el objetivo de reducir el riesgo de esta operación y con esto evitar pérdidas futuras, la institución financiera implementa metodologías y políticas que lleven a una gestión adecuada de los riesgos.

El *credit socoring* permite evaluar de manera automática el riesgo de crédito de un solicitante. Ya que se enfoca en el riesgo de incumplimiento del individuo o empresa, independiente de lo que ocurra con el resto de la cartera [Girault, 2007].

A pesar de la variedad de modelos de *credit scoring*, según [Girault, 2007], el juicio humano o del analista continúa siendo utilizado y afirma que tanto los métodos basados en la experiencia como los que se basan en análisis estadístico en la práctica muchas veces coexisten y se complementan.

Existe diversidad de metodologías en el campo del *credit scoring*, sin embargo, los modelos aditivos generalizados y la regresión logística serán abordados. Investigaciones realizadas por [Kraus, 2014] y [Liu and Chuck, 2009] evidencian el excelente desempeño de Modelos aditivos generalizados frente a modelos de credit score tradicionales como la regresión logística.

En el presente estudio se busca contrastar, si el hacer uso de modelos aditivos generalizados logra un mejor desempeño que utilizar la metodología de regresión logística en la construcción de un modelo de *credit scoring* para una cartera de crédito en una entidad financiera emergente, comparando estadísticamente cada modelo. Además, permitirá contrastar si el conjunto final de variables seleccionadas que permitirán perfilar a los clientes es el mismo o no.

Se tienen en cuenta las siguientes consideraciones: las filas de la base de datos utilizada para la construcción y validación de los modelos estadísticos corresponden a los registros u operaciones y las columnas por su parte a variables o atributos observados a cada registro. Además, la variable dependiente corresponde a una variable binaria conformada por las clases Bueno y Malo cuya definición ha sido definida previamente por la entidad financiera.

2.1 Naturaleza de los modelos de respuesta cualitativa

Los modelos que son objeto de estudio tienen como objetivo principal estimar la probabilidad de incumplimiento de un cliente al momento de la concesión del crédito; para lo cual se ha definido una variable binaria conformada por las clases Bueno y Malo, definiendo Bueno y Malo de la siguiente manera:

- **BUENO:** Buen pagador, cuya definición se basará en la mora del producto.
- **MALO:** Mal pagador, complementar a la definición del bueno.

Será representada con la letra **Y**, que es la variable dependiente del modelo. Se la redefinirá en forma de una variable dicotómica y toma los siguientes valores:

$$Y = \begin{cases} 1 : \text{Si el cliente de crédito es definido como BUENO} \\ 0 : \text{Si el cliente de crédito es definido como MALO} \end{cases}$$

La definición de Bueno y Malo para la variable dependiente **Y** fue definida previamente por la Institución Financiera.

Los modelos propuestos buscan estimar la probabilidad de que la variable Y tome valores en el conjunto $\{0, 1\}$, a partir del conjunto de las variables independientes o predictoras X y se obtienen de información histórica relacionada con aspectos sociodemográficos, crediticios, de buró, etc. Las mismas que son variables tanto cualitativas como cuantitativas.

2.2 Medidas de separación (Variables cuantitativas)

En la construcción de los modelos scoring, ya sea con el uso de regresión logística o modelos aditivos generalizados existe una gran cantidad de información (variables explicativas), sin embargo, es importante conocer el tipo y calidad de información disponible, ya que podrían existir variables que influyan poco o nada sobre la variable dependiente. En el caso de los modelos scoring que buscan predecir la probabilidad de incumplimiento de un cliente en el momento de la concesión del crédito, es necesario que la información empleada permita identificar acertadamente, las características de los clientes que son nominados como Buenos y de los clientes nominados como Malos.

El problema al momento de analizar y procesar la información proporcionada por la institución para la construcción de los modelos hace necesario el uso de las medidas de separación o divergencia que indican que tanto se diferencian (divergen) las distribuciones de clientes Buenos y Malos para cada variable explicativa y cuyo propósito es conocer el poder predictivo de cada variable, reducir la dimensionalidad del conjunto y así seleccionar aquellas variables con un mayor poder predictivo (variables que influyen fuertemente sobre la variable dependiente).

Para realizar el filtrado de las variables cuantitativas más importantes se hace uso de la Prueba de Kolmogórov-Smirnov.

2.2.1 Prueba de Kolmogórov-Smirnov para dos muestras (KS).

La prueba de Kolmogórov-Smirnov (1933) o Test *KS* para dos muestras aleatorias según [Arnold and Emerson, 2011], es una prueba de bondad de ajuste del tipo no paramétrico ya que no necesita realizar suposiciones apriori sobre la distribución de los datos y contrasta la siguiente hipótesis: Dos muestras aleatorias provienen de distribuciones continuas idénticas.
A continuación, se describe el Test *KS* para dos muestras aleatorias. Consideremos:

- $x_1, x_2, ..., x_{N_1}$ una muestra aleatoria de tamaño N_1 de una variable aleatoria continua

X con función de distribución F_1.

- $y_1, y_2, ..., y_{N_2}$ una muestra aleatoria de tamaño N_2 de una variable aleatoria continua Y con función de distribución F_2.

Con lo cual, se contrastan las siguientes hipótesis:

$$\begin{cases} H_0 : F_1(x) = F_2(x) \quad \forall x \\ H_1 : F_1(x) \neq F_2(x) \end{cases} \tag{2.1}$$

El estadístico *KS* utilizado para rechazar o no la hipótesis nula (H_0) hace uso de la función de distribución acumulada de X y de Y. El estadístico de prueba viene dado por:

$$KS = max_x \left| ecdf_1(x) - ecdf_2(x) \right|, \tag{2.2}$$

donde $ecdf$ representa la función de distribución empírica de una variable aleatoria continua. Así, podemos decir que el estadístico *KS* es la distancia máxima entre la función de distribución empírica de X y de Y y su valor varía entre 0 y 1, donde valores cercanos a 1 indican que las distribuciones difieren, mientras que valores cercanos a 0 indican que las distribuciones son idénticas. Así, se justifica el uso del estadístico *KS* como medida de divergencia entre las distribuciones de dos variables aleatorias continuas.

La hipótesis nula (H_0) mensionada anteriormente se rechaza siempre y cuando el estadístico *KS* sea mayor a su valor crítico KS_α, para un nivel de significancia α dado.

El Test *KS* será utilizado en este proyecto con la finalidad de realizar un análisis exhaustivo del comportamiento de las distribuciones de los individuos etiquetados como $Bueno$ y $Malo$; y seleccionar únicamente las variables que generen la mayor divergencia entre ellos. El proceso consiste en comparar las distribuciones empíricas del grupo $Bueno$ y $Malo$; y seleccionar las variables con el **estadístico *KS*** mayor a un valor específico diferente de 0.

2.3 Medidas de asociación (Variables cualitativas)

Las variables categóricas consideradas a formar parte del modelo son seleccionadas a partir de su poder predictivo, las mismas que se emplean en la construcción un modelo de predicción óptimo. Para lograr esto, se hace uso de las denominadas medidas de asociación, que permiten realizar un filtrado previo y con ello, seleccionar las variables categóricas con mayor poder predictivo. Estas medidas son aplicadas sobre los atributos categóricos, sin embargo, cumplen con la misma función que las medidas de divergencia.

A continuación, se describe la medida más utilizada en la práctica.

2.3.1 Valor de información (VI).

El *valor de información* de una variable categórica en problemas de clasificación binaria $(Bueno/Malo)$, según [Finlay, 2010], es probablemente la medida de asociación más popular que permite cuantificar el poder predictivo de una variable para decidir que tan bien discrimina las clases de la variable dependiente.

El *valor de información* para una variable categórica se calcula como:

$$VI = \sum_{i=1}^{n} \left(\frac{b_i}{B} - \frac{m_i}{M} \right) \times \ln \left(\frac{{}^{b_i}/_{B}}{{}^{m_i}/_{M}} \right), \tag{2.3}$$

donde:

- n : Número de categorías en que se ha clasificado la variable categórica.
- b_i : Número de elementos etiquetados como bueno dentro de la categoría i.
- m_i : Número de elementos etiquetados como malo dentro de la categoría i.
- B : Número total de elementos etiquetados como bueno.
- M : Número total de elementos etiquetados como malo.

Los valores de información están en el rango de cero a infinito, sin embargo, los valores más comunes son los que se encuentran entre 0 y 1, y mientras más grande sea

su valor, más predictiva será la variable categorizada, es decir, aportará significativamente al rendimiento del modelo.

En el presente estudio se elegirán las variables categóricas con el mayor valor de información, teniendo en cuenta que las variables con valores superiores a $0,5$ serán revisadas ya que, según [Siddiqi, 2006] puede existir sobreestimación. Además, propone la siguiente regla general para valores del *VI* aceptados:

Tabla 2.1: Valor de Información (VI)

Valor Informativo	Poder Preditivo
Menor a 0,02	No predictivo
0,02 a 0,1	Débil
0,1 a 0,3	Medio
Mayor a 0,3	Fuerte

Fuente: [Siddiqi, 2006]
Elaboración: Propia

2.4 Modelo de regresión logística múltiple - logit

2.4.1 Introducción

La Estadística es la ciencia de los datos, que tiene como objetivo fundamental la recopilación de información, clasificación, síntesis, organización y análisis de un conjunto de datos, que permiten describir la relación entre las variables y facilitan la explicación y predicción de fenómenos reales. Esto se logra con la ayuda de ecuaciones, funciones o fórmulas matemáticas que relacionen dichas variables, es decir, construir un modelo estadístico capaz de explicar la realidad.

La modelación estadística permite explicar la relación en caso de existir entre una **variable dependiente** Y, y un conjunto de p **variables independientes (explicativas)** $X_1, X_2, \ldots, X_p$. Esta relación puede expresarse formalmente como

$$Y = f\left(X_1, X_2, \ldots, X_p\right) + \epsilon, \tag{2.4}$$

donde f es una función desconocida fija de $X_1, X_2, \ldots, X_p$ y ϵ es un término de error aleatorio, distribuido de forma idéntica e independiente de las variables explicativas y tiene media igual a cero.

Con ayuda de determinadas técnicas de aprendizaje estadístico se estima una función f que satisfaga la igualdad (2.4) en un conjunto datos en particular, esta función también se puede usar para predecir el valor de Y para diferentes $X_1, X_2, \ldots, X_p$.

La mayoría de los métodos o técnicas que permiten aproximar la función f, según [James et al., 2014] pueden clasificarse como: **métodos paramétricos** y **métodos no paramétricos**.

2.4.2 Modelos Paramétricos.

Los modelos paramétricos como el modelo lineal y el modelo lineal generalizado, asumen que la función f tiene una forma funcional específica. Los modelos paramétricos

siempre se estiman con un enfoque basado en modelos de dos pasos.

1. Se realiza un supuesto sobre la forma funcional de f. La suposición más simple sobre f es que sea lineal y se lo expresa de la siguiente forma:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \tag{2.5}$$

   donde f, no es más que una combinación lineal entre las variables explicativas y los parámetros β. Al realizar esta suposición, se plantean restricciones estrictas sobre la forma de f y el problema de estimar la función f se simplifica considerablemente. El espacio de búsqueda de funciones se reduce y en lugar de estimar una función p-dimensional $f(X_1, X_2, \ldots, X_p)$ arbitraria, solo se estiman los $p + 1$ coeficientes $\beta_0, \beta_1, \ldots, \beta_p$.

2. Luego de seleccionar una forma funcional para f, se necesita un procedimiento adecuado que utilice el conjunto de entrenamiento y permita ajustar o entrenar dicha función. Para el ajuste del modelo lineal (2.5), se necesita encontrar (estimar) los valores de los $p + 1$ parámetros $\beta_0, \beta_1, \ldots, \beta_p$ de modo que

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \tag{2.6}$$

   Uno de los métodos que permite estimar los $p + 1$ parámetros se conoce como mínimos cuadrados.

El enfoque descrito permite asumir una forma funcional específica de f y simplifica el problema de ajustar una función p-dimensional a uno mucho más sencillo, el cual consiste en estimar un conjunto de parámetros, como $\beta_0, \beta_1, \ldots, \beta_p$ en (2.5). Motivo por el cual, se lo conoce como **paramétrico**.

El **Modelo Lineal Generalizado** es un algoritmo de aprendizaje supervisado que generaliza el modelo lineal clásico, de manera que la variable dependiente Y está relacionada linealmente con las covariables mediante una determinada función de enlace (Función Link). Además, el modelo permite que la variable Y se distribuya de manera diferente a una distribución normal (binomial, poisson, gamma, entre otras). Por lo general se utilizan cuando las variables bajo estudio incluyen datos categóricos.

Estos modelos forman parte del enfoque paramétrico e incluyen: la regresión lineal, modelos logísticos, etc. A continuación, se estudiará a detalle el **Modelo de regresión logística múltiple**.

2.4.3 Modelo de regresión logística múltiple - logit

El modelo logit es un algoritmo de aprendizaje supervisado que forma parte de los métodos paramétricos de regresión más utilizados para predecir la probabilidad de una variable dependiente categórica (o cualitativa) mediante una gama de variables explicativas o independientes que pueden ser cualitativas o cuantitativas. En este caso la variable dependente del Modelo de regresión logística múltiple es binaria $(Bueno/Malo)$ opciones que para fines del estudio las observaciones con etiqueta $Bueno$ toman el valor de 1 y 0 para las observaciones con etiqueta $Malo$.

Se procede a describir el modelo logit. Consideremos:

- n : Número de individuos en una muestra aleatoria.
- p : Número de varibles explicativas o independientes.
- $X = (X_1, X_2, ..., X_p)$: Conjunto de p variables independientes.
- $X_i = (1, x_{i1}, x_{i2}, \dots x_{ip})$: Vector de características del individuo i, donde x_{ij} es el valor de la variable $j = 1, 2, \dots, p$ en el individuo $i = 1, 2, \dots, n$.
- $Y = (y_1, y_2, ..., y_n)$: La variable dependiente, donde y_i es el valor de la variable Y en el individuo i. El valor y_i, representa lo siguiente:

$$y_i = \begin{cases} 1 : \text{Si el individuo } i \text{ es etiquetado como Bueno.} \\ 0 : \text{Si el individuo } i \text{ es etiquetado como Malo.} \end{cases} \tag{2.7}$$

- $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$: Vector de $p + 1$ parámetros (constantes) desconocidas que permitirán relacionar las variables independientes X, con la variable dependiente Y. Las mismas que deben ser estimadas.

La función de distribución logística es base del modelo logit, está definida de la siguiente manera:

$$\pi_i = Pr\left(y_i = 1|X_i\right) = \frac{1}{1 + exp\left(-\tau_i\right)}, \quad -\infty < \tau_i < \infty, \quad i = 1, 2, \ldots, n \tag{2.8}$$

con,

$$\tau_i = X_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Donde:

- π_i : Es la función permite encontrar la probabilidad de que $y_i = 1$, el individuo i sea etiquetado como $Bueno$, tomando en cuenta las características X_i que posee. El rango de esta función está en el intervalo $[0, 1]$

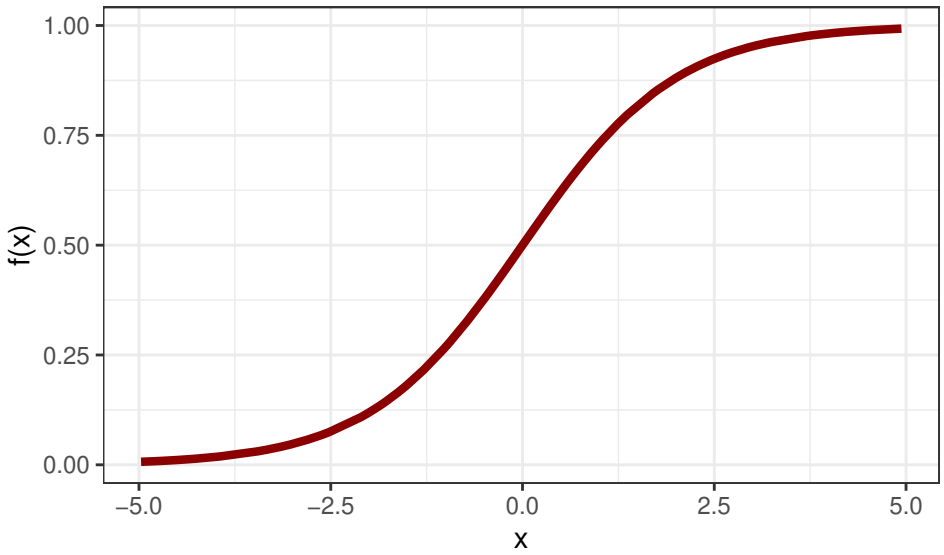
El modelo tiene como objetivo estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ que mejor se ajusten a la ecuación (2.8).

Despejando τ_i de (2.8) se obtiene la igualdad conocida como modelo $logit$. Se la presenta a continuación:

$$\tau_i = logit\left(\pi_i\right) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{2.9}$$

A la ecuación (2.9) se la conoce como razón de probabilidades $(odds\ ratio)$. La *función logística* puede ser representada gráficamente, ver Figura 2.1:

Figura 2.1: Función logística



Fuente: Elaboración propia.

La estimación de los coeficientes β puede realizarse a partir del método de máxima verosimilitud (MV), [Gujarati and Porter, 2010].

2.4.4 Interpretación de los parámetros β.

La ventaja que posee un modelo $logit$ se encuentra en su fácil interpretación. Para ayudar en la interpretación de los coeficientes es necesario conocer el significado de $odds$. La mayoría de personas, según [Allison, P. D., 2012], entienden como probabilidad a la forma "natural"de determinar la cantidad de posibilidades de que un suceso ocurra, con valores que se encuentran entre $[0, 1]$. El $odds$ forma parte de las otras formas que permiten representar el cambio "natural"de un suceso.

El término $odds$ se define como la razón que se establece entre la probabilidad de ocurrencia de un suceso y la probabilidad de no ocurrencia del mismo. La relación que existe entre el $Odds$ y la probabilidad es:

$$Odds = \frac{Probabilidad\ de\ que\ un\ suceso\ ocurra}{Probabilidad\ de\ que\ un\ suceso\ no\ ocurra} \tag{2.10}$$

Esta relación tiene importancia en el modelo $logit$. Pues, si tomamos en cuenta la

ecuación (2.8) y la relacionamos con la ecuación (2.10) se obtiene:

$$\frac{\pi_i}{1-\pi_i} = \frac{\frac{1}{1+exp(-\tau_i)}}{\frac{exp(-\tau_i)}{1+exp(-\tau_i)}} = exp\left(\tau_i\right) \tag{2.11}$$

Este resultado se conoce como transformación $logit$ de la probabilidad π_i y la relación $\frac{\pi_i}{1-\pi_i}$ como razón de probabilidades ($Odds\ ratio$). Si tomamos el logaritmo natural de (2.11) y se reemplaza $\tau_i = X_i^T\beta$, se obtiene como resultado la ecuación (2.9):

$$logit\left(\pi_i\right) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i^T\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{2.12}$$

Es debido a esta expresión que a la regresión logística se la conoce como modelo $logit$. De manera que la interpretación de este modelo está dada por la ecuación (2.12). Por ejemplo, el parámetro β_1 mide el cambio en $logit\left(\pi_i\right)$ ocasionado por un cambio unitario en x_{i1}, mientras las demás variables permanecen constantes [Gujarati and Porter, 2010].

El $Odds\ ratio$ interpreta el modelo como el cambio estimado del logaritmo natural de las probabilidades en favor de la variable dependiente cuando cada una de las variables independientes cambia en una unidad (relaciona ambas variables). La interpretación del $Odds\ ratio$, según [Gujarati and Porter, 2010] es la siguiente:

- $Odds\ ratio > 1$: Significa que existe una relación positiva (o directa).
- $Odds\ ratio < 1$: Significa que existe una relación negativa (o inversa).
- $Odds\ ratio = 1$: Significa que no existe una relación.

2.4.5 Modelos no paramétricos

Elegir un modelo puede imponer condiciones muy diferentes a la forma funcional real de f, esta es la desventaja principal de cualquier enfoque paramétrico, en cuyo caso el modelo resultante no se ajustará bien a los datos y conducirá a una estimación errónea.

Este problema se puede resolver con el uso de los denominados modelos no paramétricos, que son modelos más flexibles que permiten más formas funcionales posibles

para f.

Los modelos no paramétricos, según [James et al., 2014], a diferencia de los modelos paramétricos, antes de ajustar la función f no realiza suposición alguna sobre la forma funcional de la misma. Este método ofrece mucha más flexibilidad que los modelos paramétricos, al permitir que la forma funcional de f se pueda asumir dentro de un rango más amplio de funciones posibles. Esto resulta interesante y es la principal ventaja, pues si se hace uso de un modelo paramétrico (Regresión Logística, por ejemplo) se tendría que realizar un Análisis Descriptivo previo para saber que función describe mejor cada una de las variables explicativas. Por ejemplo, podría ser x^2, $\log(x)$,..., etc.

Existen métodos que permiten relajar la condición de linealidad mediante el uso de funciones suaves y, son usadas en la búsqueda de f, los cuales están basados en funciones polinómicas, splines cúbicos, splines de suavizado, etc.

Anteriormente se vio que el uso de los modelos paramétricos puede implicar en una equivocada elección de la forma funcional de f, lo que conducirá a que el modelo resultante realice predicciones erróneas. Este problema se resuelve con el uso de modelos más flexibles que permitan más formas funcionales para f, pero por lo general ajustar modelos más flexibles requiere estimar más parámetros. El riesgo que se asume con estos modelos, según [James et al., 2014] es el fenómeno conocido como sobreajuste.

Esto significa que la estimación de la forma funcional de f aprende demasiado, es decir, queda muy ajustada a características específicas de los datos de entrenamiento, que son usados para estimar los parámetros, pudiendo omitir información de importancia considerable. Este escenario no se desea, ya que no se conseguirán predicciones precisas para nuevas observaciones que no se encontraron dentro del conjunto de datos de entrenamiento.

Además del problema de sobreajuste. Según [Studenmund, 2016], las variables inde-

pendientes correlacionadas pueden ser un problema con respecto a la interpretación de los coeficientes, porque no es posible aislar completamente el efecto individual de cada variable independiente. Además, las variables independientes correlacionadas afectan la varianza de las estimaciones del coeficiente, lo que lleva a pruebas de significación distorsionadas.

En la práctica, se debe aplicar un procedimiento por pasos para incluir solo variables explicativas que agreguen un poder predictivo significativo al modelo. La inclusión de variables explicativas altamente correlacionadas puede causar problemas si se intentan las interpretaciones de los efectos individuales de las variables explicativas. Al incluir variables altamente correlacionadas, tales interpretaciones deben evitarse, debido a los fenómenos multicolineales. Sin embargo, si un modelo se construye únicamente con el propósito de predicción, entonces la multicolinealidad no será motivo de preocupación.

Los métodos no paramétricos al no reducir el problema de estimar f a un pequeño número de parámetros $\beta_0, \beta_1, \ldots, \beta_p$, como sucede con los métodos paramétricos, es necesario un número mayor de observaciones (mayor al que se necesita con un enfoque paramétrico) que permiten obtener una estimación óptima de f. Estos métodos admiten variables correlacionadas dependiendo de la necesidad del modelo que se esté realizando .

Los métodos paramétricos presentan significativas ventajas sobre los métodos no paramétricos cuando estamos interesados principalmente en la inferencia, ya que estas técnicas obtienen mejores interpretaciones sobre la relación entre la variable dependiente y las variables explicativas. Por ejemplo, en el modelo lineal (2.5), no presentará mayor dificultad comprender la relación existente entre Y y las variables explicativas $f\left(X_1, X_2, \ldots, X_p\right)$.

Por otro lado, al usar modelos no paramétricos pueden conducir a estimaciones muy complicadas de f que llega a ser un problema la interpretación de los resultados y resulta difícil comprender la relación entre cualquier variable explicativa con la variable

dependiente Y.

El ***Modelo Aditivo Generalizado*** ofrece un punto medio, pues puede adaptarse a relaciones complejas y no lineales entre la variable dependiente Y y las variables explicativas $f(X_1, X_2, \ldots, X_p)$ mientras se mantiene la aditividad, haciendo tanto uso de elementos paramétricos como no paramétricos y hacer buenas predicciones, pero aún permite hacer estadísticas inferenciales, comprender y explicar la estructura subyacente de los modelos y el porqué de las predicciones que hacen. A continuación, se profundiza más sobre este modelo.

2.5 Modelo Logístico Aditivo Generalizado

2.5.1 Introducción.

El **Modelo Aditivo Generalizado**, presentado por Hastie y Tibshirani (1986,1990), es una extensión de los Modelos Lineales Generalizados (GLM) al permitir funciones no lineales en cada una de las variables explicativas, mientras se mantiene la aditividad.

El modelo lineal generalizado como se vio anteriormente expresa la esperanza condicionada de la variable dependiente como combinación lineal de las variables explicativas, es decir:

$$g(u_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Sin embargo, podría ocurrir que la relación entre la variable respuesta Y_i y las variables explicativas tengan una forma desconocida. En tal situación, la estructura del modelo toma la siguiente forma [Wood, Simon N., 2017]:

$$g(u_i) = X_i^* \theta + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + \cdots + \epsilon_i \tag{2.13}$$

donde $u_i \equiv \mathbb{E}(Y_i)$ con $Y_i \sim$ alguna distribución de la Familia Exponencial, $g()$ es la función de enlace, X_i^* es la $i - ésima$ fila de la matriz del modelo correspondiente a las variables explicativas que definen cualquier componente del modelo estrictamente paramétrico, θ es el vector de coeficientes correspondiente y f_j son las funciones

suaves (funciones que tienen derivadas continuas hasta cierto orden) de las variables que se modelan no paramétricamente.

En el Modelo Aditivo Generalizado, a diferencia del Modelo Lineal Generalizado el supuesto de linealidad es relajado y proporciona suficiente flexibilidad permitiendo que no existan suposiciones sobre la forma funcional de las funciones $f_j(\cdot)$ y pueda tomar en cuenta las relaciones no lineales y así incorporar componentes no paramétricos mediante splines de regresión, por ejemplo.

Sin embargo, esta flexibilidad lleva consigo dos necesidades:

- Cómo representar las funciones suaves.
- Cómo suavizar estas funciones.

[Hastie and Tibshirani, 1990], describió muchos enfoques para estimar las funciones suaves. Algunos de los métodos de suavizamiento son los splines de suavizado y splines de regresión que se diferencian de otros métodos entre los que se encuentra el suavizamiento por núcleos (Kernel Smoothing) y regresión polinomial local (LOESS), entre otros.

La diferencia entre estos dos grupos radica en la forma en la que se realiza la estimación de los Modelos Aditivos Generalizados cuando las funciones suaves f_i están completamente parametrizadas. Los splines de suavizado, como los splines de regresión cúbica y los splines de regresión de placa delgada, se pueden expresar utilizando expansiones de base y, por lo tanto, PIRLS (mínimos cuadrados reponderados iterativamente penalizados) se puede aplicar directamente a los Modelos Aditivos Generalizados con splines de regresión.

Sin embargo, el suavizamiento por núcleos, como la regresión polinomaial local, no se pueden expresar mediante la expansión de la base, por lo que el algoritmo IRLS o PIRLS no se aplica directamente en este caso. Para realizar la estimación, se puede utilizar en su lugar un algoritmo de backfitting (Breiman y Friedman, 1985). La idea del algoritmo de backfitting es ajustar los residuos parciales de forma iterativa en cada

componente aditivo del modelo medio hasta la convergencia.

De los enfoques modernos han destacado el uso de splines de regresión para la estimación de las funciones de suavizado. [Wood, Simon N., 2017] afirma que la función suave puede representarse mejor como splines de regresión.

A lo largo de esta sección se mostrará como representar la función suave desconocida mediante splines de regresión penalizados, en particular splines cúbicos de regresión (técnica de suavizamiento empleada en este trabajo) y cómo seleccionar el parámetro de suavizado para $f_j(\cdot)$ mediante REML o "Máxima probabilidad restringida", método recomendado por [Ross, 2019]. Para poder implementarlo se hace uso de la librería **'mgcv'** del Software estadístico R, que permite estimar explícitamente los coeficientes para cada término suave mediante PIRLS (con estimación de suavidad integrada).

2.5.2 Suavizado Univariante - Funciones de Base

Con el fin de representar y estimar las funciones suaves, se considera la ecuación (2.13) simplificada a una función suave como regresor, es decir, se busca una función f que satisfaga:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{2.14}$$

donde,

- y_i : Variable dependiente.
- x_i : Variable explicativa (regresor).
- f : Una función suave.
- ϵ_i : Error aleatorio idénticamente distribuido con distribución $N(0, \sigma^2)$.

Con ayuda de los métodos estadísticos descritos anteriormente, se estima f, para lograr esto, es necesario representar f de tal manera que la ecuación (2.14) se convierta en un modelo lineal. Esto se consigue escogiendo una base, definiendo el espacio de **funciones base** b_j de dimensión $q + 1$ en donde f (o una buena aproximación), esté

presente. [Wood, Simon N., 2017], define esta base mediante una combinación lineal entre algunas **funciones básicas** conocidas y un vector de parámetros desconocidos β, con lo cual f toma la siguiente forma:

$$f(x) = \sum_{j=0}^{q} b_j(x)\,\beta_j \tag{2.15}$$

Realizando la sustitución de (2.15) en (2.14), se obtiene el siguiente modelo lineal que puede estimarse fácilmente:

$$y_i = \sum_{j=0}^{q} b_j(x)\,\beta_j + \epsilon_i, \quad i = 1, 2, \ldots, n$$

2.5.3 Regresión Polinómica

La construcción de f se puede realizar mediante la técnica conocida como **regresión polinómica**, la cual hace uso de una **base polinómica**. Esto permite ajustar la función f mediante una función polinómica de grado d, el modelo que se obtiene es el siguiente:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \epsilon_i \tag{2.16}$$

En donde $1, x_i, x_i^2, \ldots, x_i^d$ son funciones básicas y, según [James et al., 2014], los parámetros β se pueden estimar fácilmente con la técnica conocida como mínimos cuadrados.

En la práctica es inusual tomar un grado d mayor que 3 o 4. Supongamos que f es un polinomio de orden 4, de modo que una base para el espacio de funciones básicas de orden menor o igual a 4, es la siguiente:

$$b_0 = 1, \quad b_1 = x_i, \quad b_2 = x_i^2, \quad b_3 = x_i^3, \quad b_4 = x_i^4$$

Así, la ecuación (2.15) se transforma a

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

y (2.14) se escribe como

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i$$

Para un grado suficientemente grande, la **regresión polinómica** genera una curva extremadamente no lineal, demasiado flexible pudiendo así adoptar comportamientos muy extraños. Esto, conlleva a sufrir de inestabilidad en los bordes. La alternativa adecuada es la estimación por splines.

2.5.4 Splines Cúbicos de Regresión

Las splines de regresión a diferencia de la regresión polinómica permiten estimar la función f dividiendo la función original en secciones y ajustando cada sección con un polinomio individual de menor grado. Cada punto que une las distintas secciones se conoce como “nodo”.

La base que permite estimar f está formada por splines y se la conoce como **B-spline**. Una B-Spline de orden d es un polinomio de grado $d-1$, que es continuo hasta la derivada $d-2$ en la transición entre nodos.

Si tomamos $d = 3$ en la ecuación (2.16), se obtiene el siguiente modelo de regresión cúbico:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \tag{2.17}$$

En donde, los coeficientes β_0, β_1, β_2, β_3 son diferentes en las distintas secciones y el ajuste de y_i se realiza mediante un polinomio cúbico definido a trozos.

Por ejemplo, un polinomio cúbico definido a trozos con un solo nodo en un punto x' se

define de la forma:

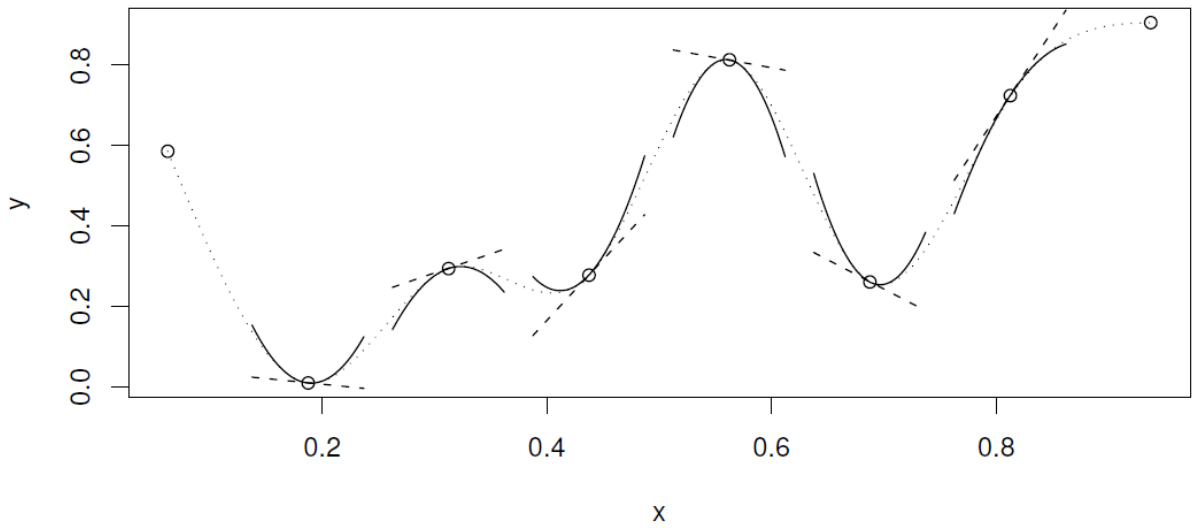
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i \quad si \quad x_i < x' \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i \quad si \quad x_i \geq x' \end{cases} \tag{2.18}$$

De modo que, (2.18) realiza el ajuste de los datos mediante dos funciones polinómicas diferentes, una para aquellas observaciones en donde $x_i < x'$ y otra para el resto de observaciones en donde $x_i \geq x'$.

Los coeficientes de ambos polinomios pueden ajustarse mediante mínimos cuadrados, en donde $\beta_{01},\ \beta_{11},\ \beta_{21},\ \beta_{31}$ y $\beta_{02},\ \beta_{12},\ \beta_{22},\ \beta_{32}$ son los coeficientes del primer y segundo polinomio respectivamente.

Gráficamente un **spline cúbico** según [Wood, Simon N., 2017], es una curva compuesta por secciones de polinomios cúbicos unidos de tal forma que la curva sea continua hasta la segunda derivada en cada nodo, ver Figura 2.2:

Figura 2.2: Spline Cúbico



Fuente: [Wood, Simon N., 2017]

En la Figura 2.2 la spline (curva punteada) tiene siete secciones en donde los puntos de unión (○) son los nodos de la spline. Cada sección tendrá distintos coeficientes, pero coincidirá el valor en los nodos con sus secciones vecinas y las primeras dos derivadas.

El uso de más nodos produce un ajuste sobre f extremadamente flexible. Si se eligen k nodos a lo largo de la curva, se termina ajustando $k+1$ polinomios cúbicos distintos. En la práctica se hace uso de los denominados **splines cúbicos** que no son más que splines de de grado igual a 3.

2.5.4.1 Splines Cúbicos - Bases

Existen algunas alternativas, pero equivalentes, de escribir una base para splines cúbicos, [Wood, Simon N., 2017] propone una de las bases más usadas, la cual hace uso de $q-2$ nodos x_i' con $i = 1, 2, \ldots, q-2$. Las funciones básicas para esta base son:

$$b_1(x) = 1, \quad b_2(x) = x, \quad b_{q+2}(x) = R(x, x_i'),$$

con

$$R(x,z) = \frac{\left[\left(z-\frac{1}{2}\right)^2 - \frac{1}{12}\right]\left[\left(x-\frac{1}{2}\right)^2 - \frac{1}{12}\right]}{4} - \frac{\left[\left(|x-z|-\frac{1}{2}\right)^4 - \frac{1}{2}\left(|x-z|-\frac{1}{2}\right)^2 + \frac{7}{240}\right]}{24}.$$

Esta base cúbica permite ajustar f de tal forma que (2.14) se transforma en un modelo lineal $y = X\beta + \epsilon$, en donde la $i-ésima$ fila de la matriz X es:

$$X_i = \left[1, x_i, R(x_i, x_1'), R(x_i, x_2'), \ldots, R\left(x_i, x_{q-2}'\right)\right]$$

Otra base muy usada es la propuesta por [James et al., 2014], la cual usa k nodos x_i' con $i = 1, 2, \ldots, k$. Cuyas funciones básicas son:

$$b_1(x) = x, \quad b_2(x) = x^2, \quad b_3(x) = x^3,$$
$$b_4(x) = h(x, x_1'), \quad b_5(x) = h(x, x_2'), \ldots, b_{k+3}(x) = h(x, x_k'),$$

con

$$h\left(x, x_i'\right) = \left(x - x_i'\right)_+^3 \begin{cases} \left(x - x_i'\right)^3, & si \quad x > x_i' \\ 0 \quad , & si \quad x \leq x_i' \end{cases} .$$

De manera que (2.14) se transforma a

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 h\left(x, x_1'\right) + \beta_5 h\left(x, x_2'\right) + \cdots + \beta_{k+2} h\left(x, x_k'\right) + \epsilon_i$$

Así, el vector de coeficientes desconocidos β puede estimarse por mínimos cuadrados en ambas bases.

2.5.5 Grado de suavizado: Splines de regresión penalizadas

El orden de la base es crucial para controlar el grado de suavizado de cualquier spline de regresión. Sin embargo, [Wood, Simon N., 2017] afirma que la elección del orden de la base no es suficiente para controlar la suavidad de la spline resultante. La base ayuda a la construcción de la función, pero no es suficiente para imponer la suavidad requerida. Las **Splines de regresión penalizadas** permiten controlar la suavidad de la función y responden a la interrogante de como suavizar esta función.

Como se vio anteriormente, el modelo se ajusta al minimizar

$$\|y - X\beta\|^2. \tag{2.19}$$

Una posibilidad para controlar la suavidad es fijar la dimensión de la base en un tamaño que sea un poco más grande de lo que se creería pueda ser necesario y controlar la suavidad del modelo añadiendo una penalización de “ondulación” al modelo, que es ajustado por mínimos cuadrados.

$$\|y - X\beta\|^2 + \lambda \int_0^1 \left[f''(x)\right]^2 dx, \tag{2.20}$$

donde λ se conoce como *parámetro de suavizado* y penaliza a una función f “ondu-

lada" al dar más o menos peso al cuadrado integrado de la segunda derivada.

El valor de λ controla la compensación entre el ajuste y la suavidad del modelo, y su elección es crucial para función resultante. Un valor de $\lambda = 0$ conduce a una estimación de spline de regresión no penalizada, es decir, crea una función que pasará directamente por cada punto de datos. Mientras que un valor de $\lambda \longrightarrow \infty$ da como resultado una estimación para f que genera una línea recta. Como la función estimada f es lineal con respecto al vector de parámetros β, la integral puede ser calculada como sigue:

$$\int_0^1 \left[f''(x)\right]^2 dx = \beta^T S \beta.$$

En donde S es una matriz de coeficientes conocidos, [Gu, 2002] muestra que

$$S_{0,0} = S_{1,1} = 0 \;\; y \;\; S_{i+2,j+2} = R\left(x_i', x_j'\right)$$

Así, el ajuste por **Splines de regresión penalizadas** se realiza minimizando

$$\|y - X\beta\|^2 + \lambda \beta^T S \beta, \tag{2.21}$$

y, el estimador de mínimos cuadrados penalizados de β dado λ viene dado por

$$\hat{\beta} = \left(X^T X + \lambda S\right)^{-1} X^T y.$$

La elección exacta de la dimensión de la base y la ubicación precisa de los nodos, tienen muy poca influencia en el ajuste del modelo, siempre que la dimensión de la base sea lo suficientemente grande como para representar f. En otras palabras, estimar el parámetro de suavizado λ resume el problema de determinar el grado de suavidad del modelo.

2.5.6 Estimación del parámetro de suavizado - REML

Si la elección del valor de λ es muy grande/pequeño los datos se suavizarán por encima/debajo. Por lo tanto, es necesario estimar un λ tal que la función estimada esté

lo más cercana a la función original f.

En general existen dos métodos que son usados, estos son: métodos de error de predicción, entre los que se encuentra la “Validación cruzada” o GCV, o métodos de probabilidad marginal basados en modelos Bayesianos/mixtos de suavizado entre los que se encuentra el método conocido como “Máxima probabilidad restringida” o REML.

En este trabajo se usará la REML, ya que según [Wood, Simon N., 2017], este método es menos propenso a los mínimos locales que los otros criterios y, por lo tanto, puede ser preferible usarla para determinar un grado apropiado de suavizado para $f_j(\cdot)$.

Como GAM tiene una interpretación bayesiana, podemos tratarla como un modelo mixto estándar separando los efectos fijos y estimando los parámetros de suavizado como parámetros de varianza.

Por tanto, la función de probabilidad restringida, dado el vector de parámetros suaves, λ, se obtiene integrando β fuera de la densidad conjunta de los datos y los coeficientes

$$l_r\left(\hat{\beta},\lambda\right) = \int f(y|\beta)\, f(\beta)\, d\beta. \tag{2.22}$$

La función de probabilidad restringida depende de λ y las estimaciones $\hat{\beta}$ (a través de la penalización), pero no de los parámetros aleatorios β. Entonces un enfoque alternativo es elegir los parámetros de suavizado que maximicen la probabilidad marginal logarítmica bayesiana:

$$v_r(\lambda) = \log l_r\left(\hat{\beta},\lambda\right) = \log \int f(y|\beta)\, f(\beta)\, d\beta. \tag{2.23}$$

Es decir, se usa (2.22) para derivar vectores de prueba para λ para una iteración PIRLS (mínimos cuadrados reponderados iterativamente penalizados) anidada:

1. Dado un vector de prueba λ, se estima β usando PIRLS.
2. Se actualiza λ maximizando la probabilidad logarítmica restringida (2.23).
3. Se repiten los pasos 1 y 2 hasta la convergencia.

2.5.7 Modelo Aditivo

Los métodos propuestos ajustan una spline de regresión a una sola variable. Los modelos score de interés tienen más de una variable. El modelo aditivo modela la variable independiente como la suma de las funciones suaves a las que están sujetas las diferentes variables explicativas. Por facilidad, sin pérdida de generalidad se trabaja con dos variables explicativas x y z. Por lo tanto, el modelo aditivo toma la siguiente estructura:

$$y_i = \alpha + f_1(x_i) + f_2(z_i) + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{2.24}$$

donde α es un parámetro de intercepción, f_1 y f_2 son funciones suaves estimadas con los métodos introducidos en las secciones anteriores y ϵ_i son el término de error aleatorio i.i.d con distribición $N(0, \sigma^2)$.

Modelar y_i como la suma de las funciones de suavizado individuales $f_1(x), f_2(z)$, en lugar de una sola función $f(x, z)$ impone una condición muy fuerte, ya que $f_1(x)+f_2(z)$ es un caso especial y restrictivo de la función suave general de ambas variables $f(x, z)$. Al modelar cada función individualmente cada predictor mantiene la capacidad interpretativa del modelo. Estimar $f(x, z)$ proporciona flexibilidad superior al modelo pero una capacidad de interpretación menor. La función suave individual es un beneficio importante del modelo aditivo.

El hecho de tener más de una función en el modelo, provoca un problema de identificabilidad, cada función solo se estima dentro de una constante aditiva. En otras palabras, una constante cualquiera podría agregarse a f_1 y sustraerse de f_2 simultáneamente, sin alterar el resultado final del modelo. Este problema se resuelve utilizando splines de regresión penalizadas, estimando los coeficientes β por mínimos cuadrados penalizados

$$\|y - X\beta\|^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta, \tag{2.25}$$

y seleccionando el parámetro de suavizado λ_i mediante Máxima probabilidad restrin-

gida (REML) como se vio anteriormente para el modelo univariado simple.

2.5.8 Modelo Logístico Aditivo Generalizado

Los GAM se pueden utilizar en situaciones donde Y es cualitativo. En los modelos $Credit\ Score$ como se vio anteriormente la variable dependiente Y toma valores cero o uno. Si relacionamos la media de la respuesta binaria $u_i(X) = Pr(y_i = 1|X_i)$ mediante un modelo de regresión lineal y la función de enlace $logit$, se construye un modelo de regresión logística (2.9) para datos binarios:

$$logit\left(u_i(X)\right) = \ln\left(\frac{u_i(X)}{1-u_i(X)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{2.26}$$

Una forma natural de extender (2.26) para permitir relaciones no lineales es usar el **modelo de regresión logística aditiva**, que reemplaza cada término lineal por una forma funcional más general y la media condicional $u_i(X)$ de una respuesta Y se relaciona con una función aditiva de los predictores a través de una función de enlace $logit$

$$logit\left(u_i(X)\right) = \ln\left(\frac{u_i(X)}{1-u_i(X)}\right) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}), \tag{2.27}$$

donde f_j es una función suave no especificada. La forma no paramétrica de estas funciones hace que el modelo sea más flexible. Sin embargo, la aditividad se conserva y esto nos permite interpretar el modelo de la misma manera que en un modelo de regresión logístico.

2.6 Estadísticos que evalúan el desempeño de los Modelos

2.6.1 Multicolinealidad

La multicolinealidad se define como la ocurrencia de una alta correlación entre dos o más variables explicativas entre sí en un modelo de regresión. Es decir, una variable explicativa puede expresarse como una combinación lineal del resto de variables

explicativas. Esta correlación es un problema ya que genera un error estándar relativamente grande, lo que implica que los coeficientes sean inestables y por lo tanto no sean precisos. Además, la presencia de multicolinealidad no permite distinguir el efecto individual de cada variable explicativa sobre la variable dependiente; lo que dificulta la interpretación del modelo y también crea problemas de sobreajuste.

Para calcular el grado de multicolinealidad entre las variables explicativas se emplea la medida conocida como *Factor de inflación de la varianza* (VIF) *y varianza generalizado* $(GVIF)$. Ambas medidas indican el grado en el que la varianza del coeficiente estimado de una variable explicativa aumenta debido a la correlación de esta variable con las demás variables explicativas del modelo.

El *Factor de inflación de la varianza* (VIF), se define como:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{2.28}$$

donde R_i^2, el coeficiente de determinación: es la proporción de la varianza que resulta de la regresión entre la variable explicativa i y el resto de variables explicativas. Esta medida se emplea para averiguar qué tan correlacionada se encuentra la variable explicativa i con el resto de variables explicativas. Un valor alto de R_i^2 indica una alta correlación.

Esta correlación es capturada por el VIF, cuando más cerca se encuentre el valor R_i^2 de 1, mayor será el valor de VIF y mayor será la multicolinealidad con la variable explicativa i.

(Fox y Weisberg, 2011), comentan que el VIF simple no se puede usar si hay variables con más de un grado de libertad, esto se da en variables categóricas con más de dos niveles o en variables polinómicas donde una variable requiere más de un coeficiente y recomiendan usar la función $GVIF$ (Factor de inflación de la varianza generalizada).

El VIF generalizado, denominado $GVIF$ se introdujo por primera vez en Monette G.

et al., (1992). El $GVIF$ es igual a:

$$GVIF_i = \frac{detR_i \times detR_{-i}}{detR}, \tag{2.29}$$

donde $detR_i$ es el determinante de la matriz de correlación para la variable i, $detR_{-i}$ es el determinante de la matriz de correlación para el resto de variables del modelo y $detR$ es el determinante de la matriz de correlación completa.

Para variables que poseen un solo coeficiente (un grado de libertad), $GVIF$ es igual a VIF. Para hacer que el $GVIF$ sea comparable en todas las dimensiones, (Fox y Monette, 1992) también sugirió usar $GVIF^{\left(\frac{1}{(2 \cdot DF)}\right)}$, donde DF (grados de libertad) es el número de coeficientes en el subconjunto. Esto reduce el $GVIF$ a una medida lineal. Es análogo a sacar la raíz cuadrada del VIF habitual. Los valores de VIF y $GVIF$ superiores a 10 correspondientes a una correlación múltiple de $0,95$ indican presencia de multicolinealidad [Gujarati and Porter, 2010].

2.6.2 Estadístico de Kolmogórov – Smirnov (KS)

El ***Estadístico KS*** permite medir cuan distintas son las distribuciones de acumulación empíricas de la probabilidad de incumplimiento estimada para clientes Buenos y Malos. El estadístico se crea a partir del ***Test KS*** descrito a detalle en la sección (2.2.1), no solo se lo emplea en la selección de variables cuantitativas con mayor poder predictivo, también es utilizado para medir la capacidad de clasificación del modelo, sus valores oscilan entre 0 y 1. [Anderson, 2007] menciona que un modelo con un KS por debajo del $20\,\%$ debe ser cuestionado y por encima del $70\,\%$, probablemente, sea muy bueno para ser cierto. Por lo tanto, la discriminación entre clientes Buenos y Malos es alto cuando más cercano a 1 se encuentre el valor KS. Es decir, el poder predictivo del modelo es mayor.

2.6.3 Área bajo la curva ROC

La Curva ROC (Receiver Operating Characteristic) es una representación gráfica de probabilidad que muestra el rendimiento de un modelo de clasificación evaluando la

capacidad que tiene para discriminar entre Buenos y Malos. Esta gráfica es representada mediante la relación entre:

- **Sensibilidad:** Razón de verdaderos positivos. Es decir, clientes Buenos que fueron clasificados como Buenos en el modelo.
- **1-Especificidad:** Razón de falsos positivos. Es decir, clientes Malos que el modelo clasificó como clientes Buenos.

Esto se consigue haciendo recorrer el umbral de clasificación (distintos puntos de corte) que maximiza la Sensibilidad, al mismo tiempo que minimiza el complemento de la especificidad (1-Especificidad). Este punto de corte toma valores en el intervalo $[0, 1]$.

Haciendo uso de la cura ROC se calcula uno de los índices que miden el rendimiento de un modelo de clasificación (capacidad de clasificar correctamente a los clientes) conocido como AUC (Area under the curve), que no es más que el área bajo la curva ROC.

El AUC o área bajo la curva ROC puede tomar valores entre 0 y 1. [Anderson, 2007] indica que un valor de 0 implica que las clasificaciones son erróneas en su totalidad, lo que significa que se tiene la peor medida de separabilidad. Un valor de $0{,}5$ indica que el modelo es igual a hacer una clasificación aleatoria, es decir, no tiene capacidad de separación de clases en absoluto y un valor de 1 indicaría predicciones perfectamente correctas. El valor AUC debe estar sobre $0{,}5$ y para modelos de clasificación, se consideran adecuados valores que sean superiores a $0{,}7$ [Siddiqi, 2006].

2.6.4 Coeficiente de GINI

El coeficiente de ***Gini*** es una métrica de precisión empleada para medir que tan bien un modelo $credit\ scoring$ logra distinguir a clientes Buenos y Malos. Proporciona un valor único que representa el poder predictivo sobre todo el rango de probabilidad pronosticada. Se lo calcula mediante la siguiente expresión:

$$GINI = 1 - \sum_{i=1}^{n} \left[P_b\left(i+1\right) - P_b\left(i\right)\right]\left[P_m\left(i+1\right) + P_m\left(i\right)\right], \tag{2.30}$$

donde:

- $\boldsymbol{n}$**:** Número de intervalos.
- $\boldsymbol{P_b}$**:** Proporción de buenos hasta el intervalo i.
- $\boldsymbol{P_m}$**:** Proporción de malos hasta el intervalo i.

El coeficiente de Gini toma valores en el intervalo $[0, 1]$, donde 0 indica que el modelo no logra discriminar entre las clases Buenos y Malos, mientras que, un valor igual a 1 significa que el modelo discrimina perfectamente a clientes Buenos y Malos [Anderson, 2007].

Existe una relación entre el coeficiente de $Gini$ y el AUC. Esta relación se representa mediante la siguiente fórmula:

$$GINI = 2 \cdot AUROC - 1, \tag{2.31}$$

esta relación indica que el coeficiente de Gini es igual a 2 veces el AUC (Área bajo la curva ROC) y la recta $y = x$. Se considera que para modelos de comportamiento, es posible obtener un coeficiente de Gini de más del $80\,\%$, mientras que un valor por debajo del $60\,\%$ podría generar sospechas [Anderson, 2007].

2.6.5 Matriz de confusión

La matriz de confusión (Kohavi y Provost, 1998), es una herramienta que permite evaluar la precisión y exactitud de un modelo de clasificación. Compara los valores reales con los valores pronosticados por el modelo para la variable objetivo.

Por ejemplo, para una matriz de un problema de clasificación binaria. Ver Tabla 2.2:

- La variable Objetivo tiene dos valores: Bueno (1) o Malo (0).

- Cada fila representa los valores pronosticados para la variable objetivo.
- Cada columna representa los valores reales para la variable objetivo.

Tabla 2.2: Matriz de Confusión

Pronóstico \ Real	Bueno	Malo
Bueno	VP	FP
Malo	FN	VN

Fuente: Elaboración propia.

En donde cada entrada de la matriz de confusión tiene los siguientes significados:

- **Verdadero Positivo (VP):** Si un cliente Bueno es clasificado como Bueno.
- **Falso Negativo (FN):** Si un cliente Bueno es clasificado como Malo.
- **Falso Positivo (FP):** Si un cliente Malo es clasificado como Bueno.
- **Verdadero Negativo (VN):** Si un cliente Malo es clasificado como Malo.

De la Tabla 2.2 se definen las siguientes métricas asociadas más importantes:

Tabla 2.3: Métricas de desempeño

Métrica	Definición	Fórmula
Precisión	Porcentaje de predicciones correctas frente al total.	$\frac{VP+VN}{TP+TN}$
Sensibilidad	Porcentaje de Buenos clasificados correctamente frente al total de Buenos.	$\frac{VP}{VP+FN}$
Especifidad	Porcentaje de Malos clasificados correctamente frente al total de Malos.	$\frac{VN}{FP+VN}$
Valor de predicción positivo	Porcentaje de Buenos clasificados correctamente frente al total de Buenos predichos.	$\frac{VP}{VP+FP}$
Valor de predicción negativo	Porcentaje de Malos clasificados correctamente frente al total de Malos predichos.	$\frac{VN}{FN+VN}$

Fuente: Elaboración propia.

2.6.6 Tablas de desempeño

Entre las herramientas empleadas para medir que tan bien clasifica correctamente el modelo a clientes Buenos y Malos son las tablas de desempeño o rendimiento, las mismas que por cada decil de Buenos estimados permite visualizar la calidad de discriminación que realiza el modelo. En general la probabilidad estimada se divide en 10 intervalos. Se analiza el número y porcentaje de clientes, número y porcentaje de clientes Buenos y Malos por cada intervalo.

La estructura de una tabla de desempeño, se divide en las siguientes secciones:

- **Probabilidad Buen Pagador:** Esta sección posee diez intervalos (abierto a la izquierda y cerrado a la derecha), donde el límite inferior es el valor mínimo y el límite superior el valor máximo de la probabilidad pronosticada.
- **Clientes Totales:** En esta sección se presentan por cada intervalo el número total de clientes (Num), el porcentaje de clientes (Porc); y por último el porcentaje acumulado (PorcAcum).
- **Clientes Buenos:** En esta sección se presentan por cada intervalo el número total de clientes Buenos (NumB), el porcentaje de clientes Buenos (PorcB); y por último el porcentaje acumulado (PorcAcumB).
- **Clientes Malos:** En esta sección se presentan por cada intervalo el número total de clientes Malos (NumM), el porcentaje de clientes Buenos (PorcM); y por último el porcentaje acumulado (PorcAcumM).
- **Tasa: Buenos/Malos:** En este campo se calcula el porcentaje de clientes Buenos y Malos respecto al total de clientes por cada intervalo de probabilidad (TasaBuenos y TasaMalos).

Un modelo credit scoring presenta un buen desempeño cuando:

- El porcentaje de clientes Buenos por cada intervalo (RazonBuenos) crece cuando la probabilidad aumenta, mientras que el porcentaje de clientes Malos por cada intervalo (RazonMalos) decrece cuando la probabilidad aumenta.

- En los deciles más altos se concentran porcentajes significativos de clientes Buenos.

Capítulo 3

Metodología Analítica

En este capítulo se muestra la metodología empleada en la construcción de un modelo $Credit\ Score$ para una cartera de crédito, lo que permite la correcta aplicación de los algoritmos de regresión logística y regresión logística aditiva generalizada.

Un modelo $Credit\ Score$ se puede aplicar según la etapa del ciclo de vida de un crédito. La Literatura señala que los modelos scoring se clasifican en dos tipos:

- **Scoring de originación:** Se utiliza en la etapa en la que se otorga la operación de crédito por primera vez en una Institución Financiera. Es decir, se aprueba o rechaza las solicitudes de nuevos créditos.
- **Scoring de comportamiento:** Se estructura para ser utilizado en la etapa de seguimiento del crédito. Es decir, permite dar seguimiento a los clientes que cuentan con historial de crédito en la institución.

Este trabajo se enfoca en el desarrollo de modelos $Credit\ Score$ de comportamiento. En donde los modelos determinarán la probabilidad de incumplimiento de los clientes que cuentan con historial crediticio. Para lo cual partiremos del conjunto de datos inicial, describiendo su forma de generación y la información que se dispone, realizaremos análisis y tratamiento de las variables explicativas, analizaremos y seleccionaremos las variables explicativas especificando criterios de selección empleados. Por último, se construye y presenta los resultados del modelo logístico aditivo generalizado con los cuales se realiza la validación estadística del mismo.

3.1 Descripción de la Base de datos

Para la creación de los modelos estadísticos mencionados anteriormente, se dispone de una Base de datos de una cartera de créditos concedidos, proporcionada por una institución financiera de un país emergente. La Base de datos original consta de $127{,}413$ observaciones referidas a las operaciones realizadas por los individuos y 61 variables explicativas que contienen las características de los clientes que permitirán discriminarlos como $Buenos/Malos$. Estas observaciones fueron tomadas entre junio de 2016 y julio de 2017, que son las fechas de observación de la operación del cliente y son consideradas para la modelización.

La información disponible de las variables explicativas proviene de tres distintas fuentes:

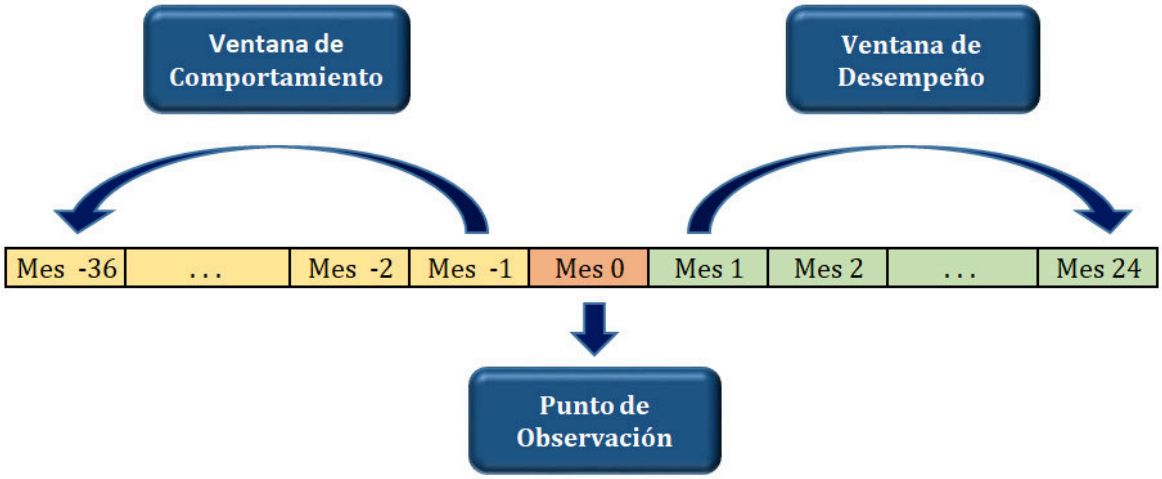
- **Información demográfica:** Son las variables propias de la solicitud de crédito y corresponden a las características sociales, económicas y demográficas.
- **Información interna:** Son las variables del comportamiento interno, describren el historial crediticio del cliente.
- **Información externa:** Son las variables que provienen de la Central de riesgos (Buró de crédito), que no son más que los datos crediticos (comportamiento) del cliente en instituciones externas.

Además, para el desarrollo de un modelo $Credit\ Score$ es necesario comprender como se realizó la elección de una muestra previa de clientes y creación de variables explicativas involucradas con su respectiva información crediticia. En la Figura 3.1 se observan dos periodos y un punto, conocidos como:

- **Ventana de Comportamiento:** Corresponde al primer periodo y sirve como punto de referencia para evaluar a un cliente como bueno o malo. Es aquí donde se construyen las variables independientes asociadas al historial crediticio del cliente. La Superintendencia de Bancos y Seguros de Ecuador dispone que este periodo no puede ser mayor a 36 meses.

- **Punto de Observación:** También conocido como mes de observación, es en donde se generan las variables socio-demográficas y se selecciona la muestra necesaria para la creación de los modelos estadísticos (Junio de 2016 a Julio de 2017).
- **Ventana de Desempeño:** Constituye los meses posteriores al punto de observación, generalmente se extiende de 12 a 24 meses y en el que se observa el comportamiento de pago de un cliente durante este periodo. Esta información es utilizada para crear la variable dependiente Y, esta permite definir a un cliente como $Bueno$ o $Malo$.

Figura 3.1: Periodos para el desarrollo de un modelo $Score$



Fuente: Elaboración propia.

Estos periodos son necesarios ya que un modelo $Credit\ Score$ se construye con el supuesto de que el comportamiento de clientes pasados reflejará comportamientos futuros y para predecir este comportamiento futuro se debe definir correctamente un cliente $Bueno$ o $Malo$.

Para este estudio la $Variable\_Dependiente$ depende de la mora del cliente y fue previamente definida por la Institución Financiera, así como también las variables explicativas fueron construidas en el periodo correspondiente, las cuales se encuentran descritas a detalle en el ANEXO 1. Teniendo en cuenta estas consideraciones, la variable dependiente cuenta con tres categorías: **Bueno**, **Malo**, **Indeterminado**, la cual se encuentra distribuida como sigue (Ver Tabla 3.1):

Tabla 3.1: Distribución de la Variable Dependiente

Y	Descripción	Clientes	Porcentaje
0	**Malo**	30.396	23,86 %
1	**Bueno**	88.228	69,25 %
2	**Indeterminado**	8.789	6,90 %
	Total	127.413	100,00 %

Fuente: Elaboración propia.

En la Tabla 3.1 podemos observar el número de clientes y su respectivo porcentaje por categoría de la variable dependiente Y, en donde el porcentaje de $Buenos$ es mayor al porcentaje de $Malos$. Su proporción es de $2,9\,\%$, es decir por cada cliente que sea considerado $Malo$, existen alrededor de 3 clientes que son considerados $Buenos$. Esto es lógico, ya que si sucediera lo contrario no sería rentable y sostenible en el tiempo para la Institución Financiera.

Cabe mencionar que a los clientes considerados como $Indeterminados$ se los excluye de la estimación de los modelos, para luego tomarlos en cuenta en la validación de los mismos. Y así, poder distinguir de manera clara a los individuos $Indeterminados$ como $Buenos$ o $Malos$ observando su comportamiento mediante otro análisis de los resultados. De esta manera, la variable dependiente a pronosticar cuenta con las categorías $Bueno/Malo$. A continuación, se procede a realizar el análisis exploratorio y tratamiento de los datos.

3.2 Análisis exploratorio y tratamiento de datos

La Base de datos proporcionada cuenta con información que como se mencionó anteriormente proviene de distintas fuentes y que hacen referencia a información: socioeconómica, sociodemográfica, comportamiento crediticio, productos de crédito en otras instituciones, comportamiento de mora, etc. Información valiosa para la estimación óptima de los modelos. Sin embargo, también existe información que se cataloga como irrelevante ya que presenta algún tipo de anomalía en su estructura o presentan incoherencias, como por ejemplo variables constantes, variables con un porcentaje de valores perdidos alto, etc.

En esta sección se procede a realizar un análisis exploratorio. Es decir, se realiza un análisis estadístico descriptivo univariado de todas las variables explicativas.

3.2.1 Análisis Univariado

Dentro de los análisis a realizar a las variables independientes se encuentra el Análisis Univariado o Estadística Descriptiva que no es más que el análisis de cada una de las variables estudiadas por separado. Las técnicas a utilizar son: Medidas de Tendencia Central, Posicionamiento y Dispersión (Variables Cuantitativas) y Tablas de Frecuencias (Variables Cualitativas). En esta etapa se busca identificar anomalías en cada una de las variables.

3.2.1.1 Medidas de Tendencia Central, Posicionamiento y Dispersión (Variables Cuantitativas)

Las medidas a calcular son: porcentaje de valores perdidos, porcentaje de ceros, mínimo, percentil $25\,\%$, mediana, media, percentil $75\,\%$, máximo y desviación estándar. Por ejemplo para la variable $V16\_Amortizacion$ que es el porcentaje entre Saldo y Monto Bruto. Ver Tabla 3.2

Tabla 3.2: V16_Amortizacion

Nº	Estadístico	Valor
1	Porc_NAs	0,00 %
2	Porc_0s	0,00 %
3	Mínimo	0,00
4	Perc_25 %	0,41
5	Mediana	0,58
6	Media	0,54
7	Perc_75 %	0,71
8	Máximo	0,95
9	Desv_Est	0,20

Fuente: Elaboración propia.

El listado con los estadísticos para las variables cuantitativas restantes se presenta en el ANEXO 2.

3.2.1.2 Tablas de Frecuencias (Variables Cualitativas)

Para cada variable cualitativa analizada se construye una tabla con el número de individuos presentes en cada categoría. A esta tabla se la conoce como tabla de frecuencias. Además, se presenta el porcentaje de valores perdidos. Por ejemplo, la Tabla 3.3 que considera la variable $V43\_EsIndependiente$ que indica si el cliente es Dependiente o Independiente.

Tabla 3.3: V43_EsIndependiente

Nº	Categorías	Frecuencia	Porcentaje
1	DEP	12.985	10,19 %
2	IND	114.426	89,81 %
3	(Missing)	2	0 %
4	Total	127.413	100 %

Fuente: Elaboración propia.

Las tablas de frecuencias para el resto de variables cuantitativas son presentadas en su totalidad en el ANEXO 2.

3.2.2 Depuración de los datos

También conocido como limpieza de datos o scrubbing, permite modificar o eliminar los datos incorrectos, incompletos, que tienen un formato incorrecto o están duplicados.

Una vez que se ha realizado una exploración preliminar de los datos se procede a realizar el tratamiento de los mismos, comenzando con la depuración de los datos. Para lo cual, se considera lo siguiente:

- **Completitud:** Se da un tratamiento a los valores faltantes o perdidos tanto de las variables cualitativas como cuantitativas.

  Se hará uso del método de imputación cuando los valores perdidos tanto de las variables cualitativas como cuantitativas no sean superiores al $5\,\%$. Para las variables cualitativas se hará uso de la moda, mientras que, para las variables cuantitativas,

primero se observa si son asimétricas o no. Cuando la variable sea asimétrica se hará uso de la mediana y la media cuando la variable sea simétrica.

En caso de que alguna variable supere el $5\,\%$ de valores perdidos, se realizará un análisis para ver si los valores faltantes tienen algún significado o no. En nuestra base solo nos encontramos con el caso especial de la variable $V58\_Antiguedad\_SUNAT$ que posee el $78{,}8\,\%$ de valores faltantes, por lo que esta variable será excluida del conjunto de variables explicativas.

- **Consistencia:** Busca el cumplimiento de reglas semánticas previamente definidas sobre las variables. Por ejemplo, que la variable edad tenga un elemento negativo, o que por ejemplo variables porcentuales estén fuera del rango $(0, 1)$.
- **Variables Constantes:** En caso de existir variables constantes, estas son exlcuidas del conjunto de variables explicativas ya que el poder de discriminación entre las clases Bueno y Malo es nulo.

Una vez realizada la depuración de los datos se procede a realizar un estudio de cada variable numérica.

3.2.3 Análisis y tratamiento de valores atípicos

No existe un procedimiento único establecido para definir e identificar valores atípicos en general, esto debido a las características particulares de cada variable numérica.

Un valor atípico es un punto de datos que es significativamente diferente de los datos restantes. Hawkins (1980) define formalmente el concepto de un valor atípico como sigue: “Un valor atípico es una observación que se desvía tanto de las otras observaciones como para despertar sospechas de que fue generado por un mecanismo diferente”. Los valores atípicos también se denominan anormalidades, desviaciones, o anomalías en la literatura sobre minería de datos y estadística. Por lo tanto, es importante identificarlos antes de modelarlos y analizarlos.

Por criterio experto, se interpretan ciertas variables numéricas sin procesar y decidir si un punto de los datos es un valor atípico o no. Así, a las variables numéricas se-

leccionadas se les realiza un estudio, analizando y tratando todos los valores atípicos que se encuentren.

Los diagramas de cajas y bigotes son una representación gráfica de datos numéricos a través de sus cuartiles inferior y superior (definidos como percentiles 25 y 75). En donde se define:

$\mathbf{Q_1}$: Es el cuartil inferior.

$\mathbf{Q_3}$: Es el cuartil superior.

$(\mathbf{Q_3} - \mathbf{Q_1})$: Se denomina rango intercuartílico o IQR.

Es una forma muy simple pero efectiva de visualizar valores atípicos. Describe el comportamiento de los datos en el medio y en los extremos de las distribuciones.

Un diagrama de cajas se construye dibujando una caja entre los cuartiles superior e inferior con una línea sólida dibujada a través de la caja para ubicar la mediana. Se necesitan las siguientes cantidades (llamadas vallas) para identificar valores extremos en las colas de la distribución:

Valla interior inferior: $Q_1 - 1{,}5 \cdot IQR$.

Valla interior superior: $Q_3 + 1{,}5 \cdot IQR$

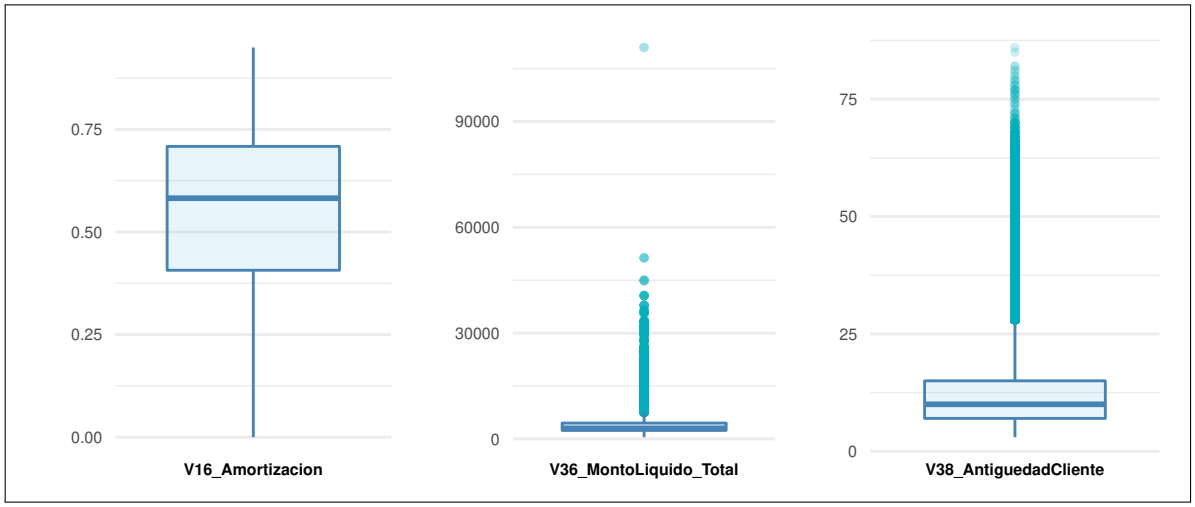
Valla exterior inferior: $Q_1 - 3 \cdot IQR$

Valla exterior superior: $Q_3 + 3 \cdot IQR$

Un punto que se encuentra más allá de una valla interior a cada lado se considera un **valor atípico leve** . Un punto que se encuentra más allá de una valla exterior se considera un **valor atípico extremo**.

A modo de ejemplo, en la Figura 3.2, se presentan los diagramas de cajas para las siguientes variables numéricas en ese orden: $V16\_Amortizacion$, $V36\_MontoLiquido\_Total$, $V38\_AntiguedadCliente$.

Figura 3.2: Diagramas de cajas y bigotes



Fuente: Elaboración propia.

Los diagramas de cajas para las variables numéricas restantes se pueden visualizar en el ANEXO 3.

3.2.3.1 Tratamiento de valores atípicos

El método empleado en el tratamiento de los valores atípicos es el conocido como **Winsorización**, que es una forma de minimizar la influencia de valores atípicos en los datos al asignar al valor atípico un peso menor, reemplazando un número específico de valores extremos con un valor de datos más pequeño. La técnica Winsorize fue introducida por primera vez por Dixon, quien la atribuyó a Charles P. Winsor.

Estadísticas clásicas, como la media y la varianza son muy susceptibles a los valores extremos. La Winsorización puede ser una forma eficaz de abordar este problema, mejorar la eficiencia estadística y aumentar la solidez de las inferencias estadísticas reduciendo el impacto de las observaciones extremas.

En este estudio se hace uso de la Winsorización del $90\,\%$. Es decir, los valores menores que el percentil $5\,\%$ se reemplazan por el valor en el percentil $5\,\%$, mientras que los valores mayores al percentil $95\,\%$ se establece igual al valor en el percentil $95\,\%$.

El método de Winsorización es una técnica estándar de la industria para tratar valores

atípicos. Funciona bien.

3.3 Selección de Muestras: Desarrollo y Validación

La Base de datos proporcionada corresponde a la muestra total seleccionada en el mes de observación, necesaria para la construcción de los modelos estadísticos. Según [Siddiqi, 2006], esta muestra debe dividirse en dos submuestras: Muestra de Desarrollo (Permite estimar los modelos) y Muestra de Validación (Empleada para validarlos).

Es usual separar una muestra representativa para el desarrollo del $70\,\%$ al $80\,\%$ del total y otra para la validación del $20\,\%$ al $30\,\%$. En este trabajo se elije la proporción $70\,\%/30\,\%$ de la muestra original mediante muestreo aleatorio simple. De tal manera que ambas muestras sean representativas para el mes de observación, logrando de esta manera una muestra representativa de buenos.

- **Muestra de Desarrollo:** Muestra utilizada para el desarrollo de los modelos y consta de 83036 individuos, que representan el $70\,\%$ de la muestra original comprendida entre Junio de 2016 Julio de 2017. Ver Tabla 3.4:

Tabla 3.4: Muestra de Desarrollo

Y	Descripcion	Clientes	Porcentaje
0	**Malo**	21.346	25,71 %
1	**Bueno**	61.690	74,29 %
	Total	83.036	100,00 %

Fuente: Elaboración propia.

- **Muestra de Validación:** Esta muestra es empleada para validar los modelos, comprende el $30\,\%$ restante de la muestra original con un total de 35588 individuos. Ver Tabla 3.5:

Tabla 3.5: Muestra de Validación

Y	Descripcion	Clientes	Porcentaje
0	**Malo**	9.050	25,43 %
1	**Bueno**	26.538	74,57 %
	Total	35.588	100,00 %

Fuente: Elaboración propia.

Por otro lado, los individuos nominados como indeterminados representan el $6,9\,\%$ de la muestra original, con 8789 individuos.

A continuación, se procede a categorizar las variables explicativas con el fin de obtener un modelo óptimo.

3.4 Categorización de variables

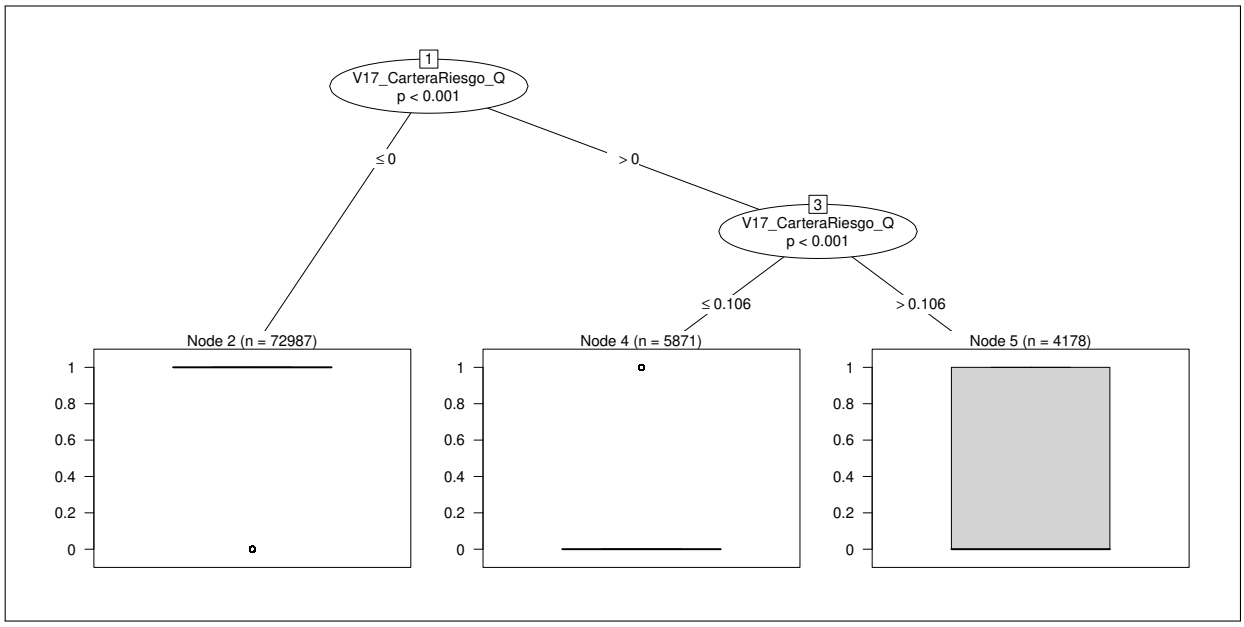
Como se mencionó anteriormente, se cuenta tanto con variables cualitativas, que son las que cuentan con distintas categorías o atributos y para el modelo serán utilizadas transformándolas variables categóricas tipo factor, así como también se cuenta con variables cuantitativas, que son aquellas variables numéricas y a las cuales se les aplicará las funciones suaves (Splines cúbicos de regresión) en el caso del Modelo Logístico Aditivo Generalizado.

En esta sección, se procede a categorizar aquellas variables que cumplan con las siguientes condiciones: variables explicativas cualitativas que posean 5 o más categorías y variables cuantitativas (continuas o nominales) que posean un valor cuya frecuencia relativa sea mayor o igual al $20\,\%$. Por otro lado, aquellas variables cuantitativas que posean menos de 5 valores únicos se las transforma a variables categóricas tipo factor.

Esta categorización se realiza mediante árboles de decisión, con grupos de al menos el $5\,\%$ de observaciones de la muestra total en cada nodo, tal como se muestra en la Figura 3.3 que considera la variable $V17\_CarteraRiesgo\_Q$, que indica el porcentaje

entre Saldo Vencido y Saldo.

Figura 3.3: Árbol de decisión - $V17\_CarteraRiesgo\_Q$



Fuente: Elaboración propia.

Como se puede observar en la Figura 3.3, el árbol establece tres categorías para la variable $V17\_CarteraRiesgo\_Q$, las cuales permiten explicar la variable dependiente Y. Los árboles de decisión restantes se pueden observar en el ANEXO 4.

3.5 Selección de las variables

En la presente sección se presentan las metodologías empleadas en la selección de variables explicativas con mayor poder predictivo, que influyan fuertemente sobre la variable dependiente Y:

- **Medidas de separación:** Como se mencionó anteriormente al estudiar las medidas de separación, la medida empleada para la correcta selección de variables explicativas cuantitativas es la prueba de Kolmogórov-Smirnov (Test KS). En la Tabla 3.6 se exponen los valores KS de todas las variables cuantitativas ordenadas de mayor a menor.

Tabla 3.6: Medida de separación KS

Nº	Variable	KS
1	V31_AtrasoMax_U6_AC	0,4913
2	V28_MaxAtraso_DC	0,4747
3	V32_AtrasoMax_U12_AC	0,4736
4	V33_AtrasoMax_U18_AC	0,4685
5	V70_Num_Calificacion_0	0,2095
6	V16_Amortizacion	0,1892
7	V38_AntiguedadCliente	0,0958
8	V61_NumeroInst_Adeuda_3	0,0803
9	V60_NumeroInst_Adeuda	0,0772
10	V62_NumeroInst_Adeuda_6	0,0681
11	V46_Edad	0,0605
12	V81_Amortizacion_comp	0,0484
13	V36_MontoLiquido_Total	0,045
14	V44_AntiguedadResidencia	0,0243
15	V91_VariacionDeudaTarjeta_U6M	0,0162
16	V68_TiempoHistorialCrediticioSF_12	0,0027
17	V90_VariacionDeudaMicrocredito_U6M	0,0018

Fuente: Elaboración propia.

De la Tabla 3.6 se puede observar que las variables explicativas que más influyen sobre la variable dependiente están relacionadas con el máximo atraso del cliente, siendo la variable $V31\_AtrasoMax\_U6\_AC$, la cual índica el máximo atraso de la persona $T6$ meses atrás hasta la fecha de corte $(T0)$ y posee el valor KS más alto con $0{,}49$ (En el ANEXO 1 se encuentra a detalle la descripción del resto de variables).

En este estudio cuando dos variables cuantitativas ingresan al modelo y tienen una correlación alta (mayor a $0{,}7$), se elige aquella que tenga el valor KS más alto, pues tendrá mayor poder predictivo y por lo tanto el modelo será más eficaz. Por lo tanto, variables cuantitativas no correlacionadas con un mayor KS forman parte del modelo final.

- **Medidas de asociación:** Para la selección de variables cualitativas se hace uso de

la medida de asociación conocida como valor de información (VI), mencionada en el capítulo 2. En la Tabla 3.7 se puede observar las variables cualitativas cuyo valor de información (VI) es mayor a $0{,}1$ ordenadas de mayor a menor. Es decir, según la Tabla 2.1 las variables con poder predictivo medio y fuerte.

Tabla 3.7: Medida de asociación VI

Nº	Variable	VI
1	V29_PromAtraso_DC	1,248029
2	V17_CarteraRiesgo_Q	0,997835
3	V25_Porc_CuotasPag	0,984001
4	V26_Porc_CuotasVenc	0,983458
5	V18_SaldoMMora_Mont	0,768008
6	V79_MoraPonderada_comp	0,422633
7	V73_PeorCalificacionCorte_Comp	0,397356
8	V74_PeorCalificacionU6M_Comp	0,383076
9	V75_PeorCalificacionU12M_Comp	0,339025
10	V76_PeorCalificacionU18M_Comp	0,31697
11	V71_Num_Calificacion_1	0,293248
12	V78_CarteraRiesgo_comp	0,208485
13	V77_CarteraRiesgoPond_U6M_Comp	0,205314
14	V82_Saldo_MB_Comp	0,20399
15	V72_Num_Calificacion_234	0,191611

Fuente: Elaboración propia.

Se dispone inicialmente de 41 variables cualitativas, de las cuales se toma en consideración las primeras 15 variables con mayor valor de información que permitirán construir el modelo final, que son las que se muestran en la Tabla 3.7 (La descripción de todas las variables se las puede observar a detalle en el ANEXO 1). El resto de variables son descartadas ya que no poseen poder predictivo alguno dentro del modelo.

En el ANEXO 5 se presenta el cálculo del valor de información de todas las variables cualitativas.

3.6 Construcción del modelo logístico aditivo generalizado

Una vez que se han seleccionado las variables explicativas, es decir, una vez que se ha realizado el análisis exploratorio y superado los estadísticos Kolmogórov Smirnov (KS) y Valor de Información (VI), se procede a construir el modelo logístico aditivo generalizado.

Cabe recalcar que a las variables cuantitativas se les aplicará las funciones suaves conocidas como splines cúbicos de regresión, permitiendo capturar la relación no lineal entre la variable explicativa y la variable dependiente Y.

Mediante el algoritmo *backward* se obtiene el modelo con mejor desempeño, que cumple con las pruebas de especificación. El proceso consiste en que a partir de un modelo inicial en el cual se incluyen aquellas variables explicativas que han sido previamente seleccionadas, la variable considerada como menos influyente por el criterio de Akaike se va eliminando en cada iteración, hasta que no existan variables a eliminar. En la Tabla 3.8 se presenta el modelo resultante $(GAM-1)$.

Tabla 3.8: Modelo logístico aditivo generalizado $(GAM-1))$

Variable	Estimate	Std,	Error z	value Pr(>|z|)
(Intercept)	4,6858934	0,0685539	68,353	<2e-16 ***
V16_Amortizacion	-3,5543249	0,0668065	-53,203	<2e-16 ***
V31_AtrasoMax_U6_AC	-0,0319800	0,0015906	-20,106	<2e-16 ***
V38_AntiguedadCliente	0,0241406	0,0014725	16,394	<2e-16 ***
V46_Edad	0,0076785	0,0009216	8,332	<2e-16 ***
V60_NumeroInst_Adeuda	-0,1594476	0,010451	-15,257	<2e-16 ***
V81_Amortizacion_comp	-0,0071248	0,0023332	-3,054	0,00226 **
V90_VariacionDeudaMicrocredito_U6M	-0,3174537	0,0318389	-9,971	<2e-16 ***
V91_VariacionDeudaTarjeta_U6M	-0,1747585	0,0307103	-5,691	1,27e-08 ***
V17_CarteraRiesgo_Q(0; 0,10607]	-1,9991640	0,0508399	-39,323	<2e-16 ***
V17_CarteraRiesgo_Q(0,10607; Inf]	-2,5143071	0,0597794	-42,060	<2e-16 ***
V18_SaldoMMora_Mont(0,0034556; 0,042912]	0,3284082	0,0466248	7,044	1,87e-12 ***
V18_SaldoMMora_Mont(0,042912; 0,059177]	0,2677341	0,0446941	5,990	2,09e-09 ***
V18_SaldoMMora_Mont(0,059177; Inf]	0,7858173	0,0539293	14,571	<2e-16 ***
V29_PromAtraso_DC(0; 1]	-0,8131556	0,0297907	-27,296	<2e-16 ***
V29_PromAtraso_DC(1; 2]	-1,2071311	0,0355982	-33,910	<2e-16 ***
V29_PromAtraso_DC(2; 4]	-1,5509610	0,0372099	-41,681	<2e-16 ***
V29_PromAtraso_DC(4; 6]	-1,9528779	0,0497364	-39,265	<2e-16 ***
V29_PromAtraso_DC(6; Inf]	-2,1985410	0,062063	-35,424	<2e-16 ***
V79_MoraPonderada_comp(0,42857; 3,05]	-0,2398773	0,0342383	-7,006	2,45e-12 ***
V79_MoraPonderada_comp(3,05; Inf]	-0,8081051	0,0273012	-29,600	<2e-16 ***
V82_Saldo_MB_Comp(0,0052869; Inf]	-0,7767923	0,0396068	-19,613	<2e-16 ***
R-sq,(adj) = 0,391 Deviance explained = 34,4 %				
-REML = 31100 Scale est, = 1 n = 83036				

Fuente: Elaboración propia.

En la Tabla 3.8, se observa que todas las variables del modelo logístico aditivo generalizado son estadísticamente significativas y los signos de los coeficientes son consistentes. Además, este modelo cumple con la prueba de multicolinealidad. Por lo tanto, se puede considerar que este modelo es adecuado.

Sin embargo, el ajuste del modelo se analizó con respecto al nivel de desviación explicado (0 a $100\,\%$; cuanto más alto, mejor) y el criterio de información de Akaike (AIC), utilizado para evaluar el ajuste y la parsimonia de cada modelo. El cual, tiene en cuenta los grados de libertad utilizados y la bondad del ajuste, de modo que los modelos más parsimoniosos tienen un AIC más bajo.

Por lo que, el siguiente paso a realizar es: observar si al suavizar las variables cuan-

titativas se logra mejorar la desviación explicada (indica que tanto el modelo logra explicar los datos). Obteniendo el siguiente modelo óptimo $GAM - 2$, ver Tabla 3.9:

Tabla 3.9: Modelo logístico aditivo generalizado final $(GAM - 2)$

Variable	Estimate	Std,	Error z	value Pr(>|z|)
(Intercept)	3,8692255	0,0741994	52,146	<2e-16 ***
V16_Amortizacion	-3,6965097	0,0682837	-54,135	<2e-16 ***
V38_AntiguedadCliente	0,0240243	0,0014957	16,063	<2e-16 ***
V46_Edad	0,0081375	0,0009347	8,706	<2e-16 ***
V60_NumeroInst_Adeuda	-0,1342892	0,0127635	-10,521	<2e-16 ***
V90_VariacionDeudaMicrocredito_U6M	-0,3045177	0,0327732	-9,292	<2e-16 ***
V91_VariacionDeudaTarjeta_U6M	-0,1570528	0,0312248	-5,030	4,91e-07 ***
V17_CarteraRiesgo_Q(0; 0,10607]	-1,9196444	0,0517271	-37,111	<2e-16 ***
V17_CarteraRiesgo_Q(0,10607; Inf]	-2,2907106	0,0592025	-38,693	<2e-16 ***
V18_SaldoMMora_Mont(0,0034556; 0,042912]	0,4360558	0,0481509	9,056	<2e-16 ***
V18_SaldoMMora_Mont(0,042912; 0,059177]	0,3783396	0,0458090	8,259	<2e-16 ***
V18_SaldoMMora_Mont(0,059177; Inf]	0,6715093	0,0529530	12,681	<2e-16 ***
V29_PromAtraso_DC(0, 1]	-0,2174578	0,0373868	-5,816	6,01e-09 ***
V29_PromAtraso_DC(1; 2]	-0,2097806	0,0495855	-4,231	2,33e-05 ***
V29_PromAtraso_DC(2; 4]	-0,2159955	0,0556717	-3,880	0,000105 ***
V29_PromAtraso_DC(4; 6]	-0,3740569	0,0669786	-5,585	2,34e-08 ***
V29_PromAtraso_DC(6; Inf]	-0,7343286	0,0741640	-9,901	<2e-16 ***
V79_MoraPonderada_comp(0,42857; 3,05]	-0,2048615	0,0347063	-5,903	3,58e-09 ***
V79_MoraPonderada_comp(3,05; Inf]	-0,7579317	0,0277241	-27,338	<2e-16 ***
V82_Saldo_MB_Comp(0,0052869; Inf]	-0,7810878	0,0401109	-19,473	<2e-16 ***

Approximate significance of smooth terms:

Variable	edf	Ref.df	Chi.sq	p-value
s(V31_AtrasoMax_U6_AC)	8,444	11	1733,92	<2e-16 ***
s(V81_Amortizacion_comp)	3,839	5	44,56	1,02e-09 ***

R-sq,(adj) = 0,405 Deviance explained = 35,9 %
-REML = 30424 Scale est, = 1 n = 83036

Fuente: Elaboración propia.

Del mismo modo que en el modelo anterior, en la Tabla 3.9 se pueden observar que todas las variables son estadísticamente significativas y cumple con la prueba de multicolinealidad. Sin embargo, en este modelo no es posible realizar un análisis de interpretación de las variables y los respectivos signos de sus coeficientes ya que son coeficientes de una forma funcional de las variables, mas no, coeficientes de las variables.

En la Tabla 3.10 se muestran la desviación explicada y el criterio de información de

Akaike que ratifican la elección del modelo $GAM - 2$, como modelo óptimo.

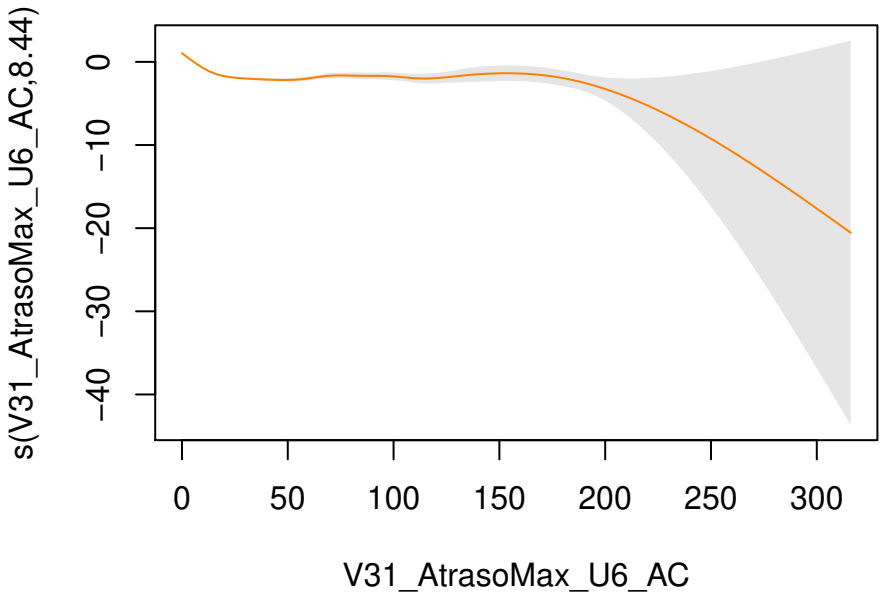
Tabla 3.10: Elección modelo óptimo

Modelo	Deviance explained	AIC
GAM-1	34,4 %	62108,17
GAM-2	35,9 %	60733,31

Fuente: Elaboración propia.

A las variables cuantitativas que ingresaron en el modelo inicial ($GAM-1$) se les aplicó las funciones suaves conocidas como splines cúbico de regresión. A continuación, se muestran los términos suaves ajustados que formaron parte del modelo óptimo ($GAM - 2$):

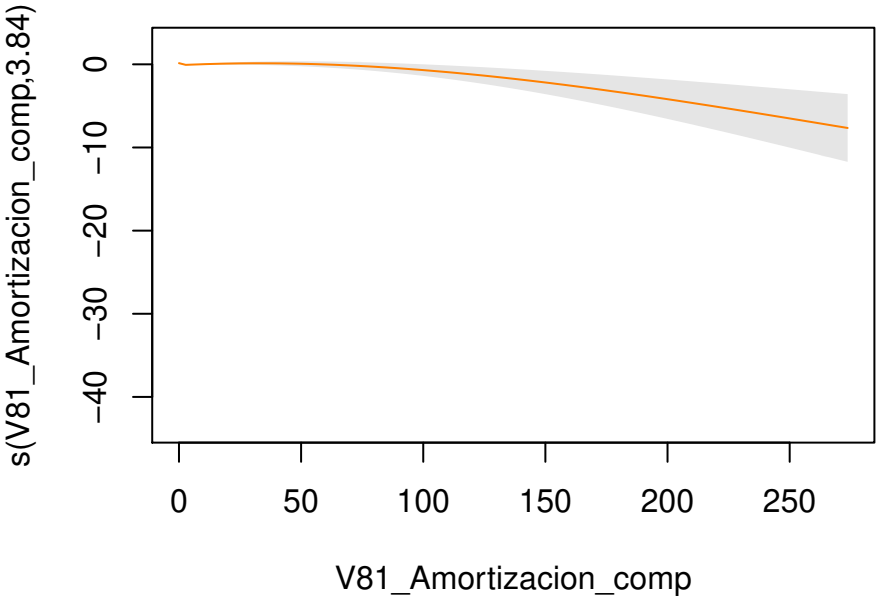
Figura 3.4: Predicción parcial para la variable $V31\_AtrasoMax\_U6\_AC$



Fuente: Elaboración propia.

En la Figura 4.6 se puede observar que la función suave ajustada tiene un comportamiento no lineal para la variable $V31\_AtrasoMax\_U6\_AC$ de modo que se justifica el uso de Modelos Aditivos Generalizados para construir un modelo adecuado, ya que se modela automáticamente esta relación no lineal que los métodos tradicionales perderán ayuda a eliminar el sesgo lineal que poseen estos métodos. Esto significa que no será necesario probar manualmente muchas transformaciones diferentes en cada variable individualmente.

Figura 3.5: Predicción parcial para la variable $V81\_Amortizacion\_comp$

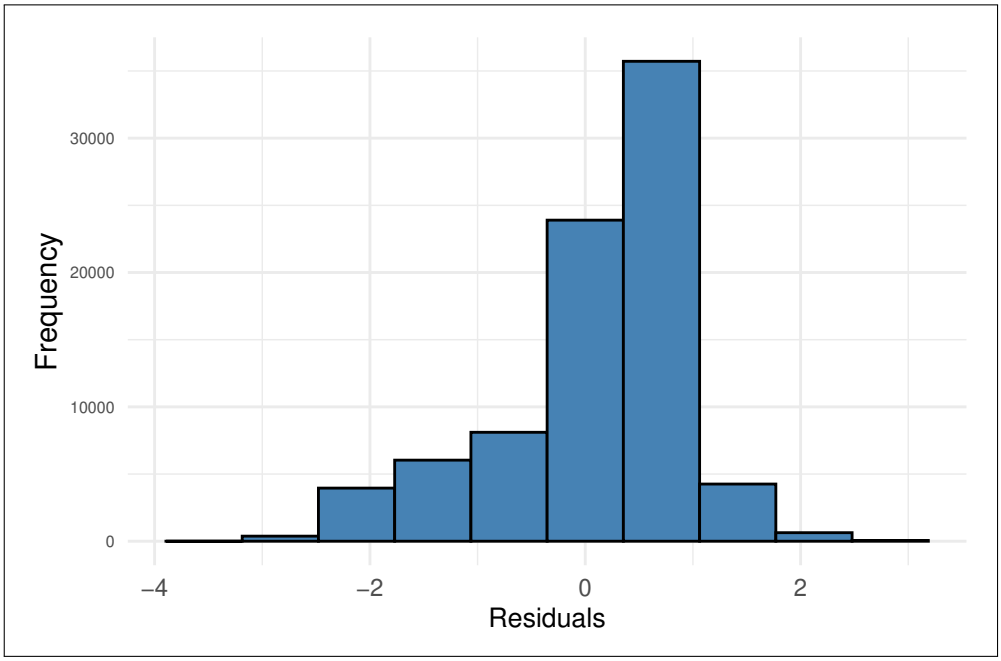


Fuente: Elaboración propia.

De la misma manera, en la Figura 3.5 se observa que existe una relación no lineal entre la variable $V81\_Amortizacion\_comp$ y la probabilidad de que un cliente sea un Buen pagador. Podemos observar que la función suave tiene un comportamiento cuadrático.

Por otro lado, se presentan los gráficos de diagnóstico para el modelo logístico aditivo generalizado óptimo $(GAM - 2)$ que permiten verificar visualmente que el modelo construido es bueno y ajusta correctamente a los datos sin presentar problemas de sobreajuste.

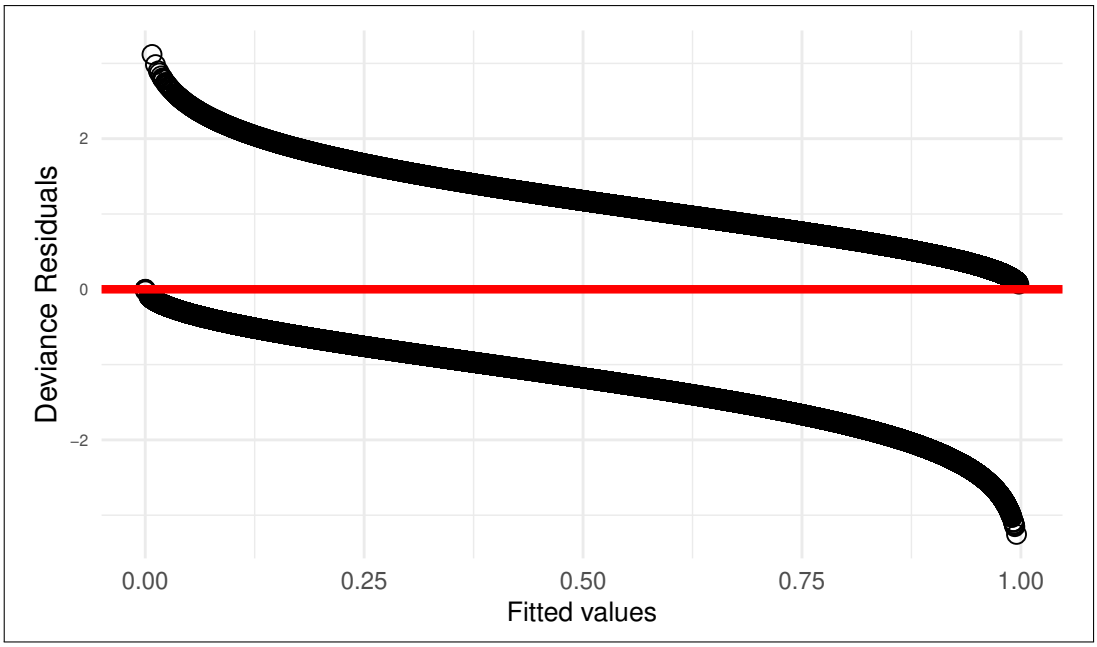
Figura 3.6: Histograma de residuos para el Modelo $(GAM - 2)$



Fuente: Elaboración propia.

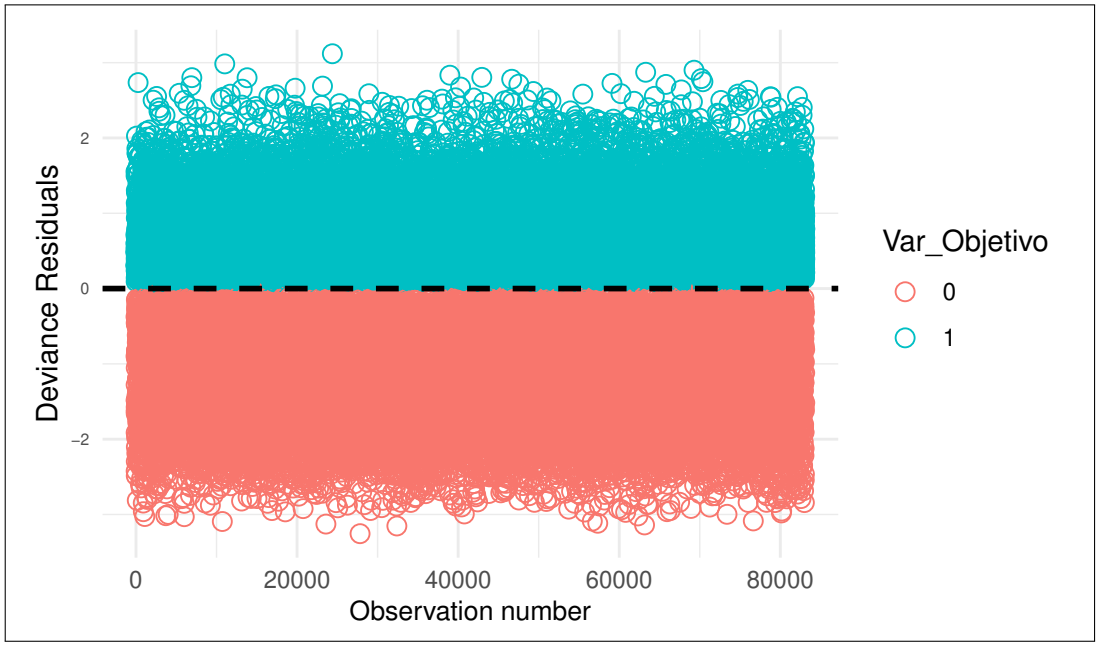
A pesar de que un modelo con respuesta binaria no supone que los residuos estén distribuidos normalmente ni que la varianza sea constante, se observa en la Figura 3.6 que el histograma de los residuos no se aleja demasiado de una distribución normal. Sin embargo, la desviación residual es útil para determinar si los puntos individuales no se ajustan bien al modelo.

Figura 3.7: Residuos de devianza frente a la probabilidad ajustada - Modelo $(GAM - 2)$



Fuente: Elaboración propia.

Figura 3.8: Residuos de Devianza frente a las obervaciones índice - Modelo $(GAM - 2)$



Fuente: Elaboración propia.

Como se puede apreciar, se han generado dos gráficos: el gráfico de los residuos de Desviación vs. los valores ajustados (Figura 3.7) y el gráfico de los residuos de Desviación vs. las observaciones índice (Figura 3.8). En donde se puede apreciar que no existen observaciones que estén muy lejanas de la mayoría de las otras observaciones; además, únicamente se debe verificar la no existencia de patrones en los residuos, por lo que el modelo es adecuado debido al poder predictivo que posee.

En la siguiente sección, se muestran algunos resultados a través de los cuales se analiza el rendimiento del modelo.

3.7 Resultados y evaluación del modelo logístico aditivo generalizado

La presente sección se centra en la evaluación estadística del modelo previamente construido. Con ayuda del conjunto de datos (Muestra de Validación) y técnicas estadísticas definidas en la sección (2.6) se evalúa la capacidad predictiva y de discriminación del modelo logístico aditivo generalizado.

- **Multicolinealidad:** Para analizar el problema de multicolinealidad , se realizó el

cálculo del factor de inflación generalizado ($GVIF$) para cada variable paramétrica del modelo $GAM-2$ y se obtuvieron los siguientes resultados:

Tabla 3.11: Factor de inflación generalizado $(GVIF)$ - Modelo $(GAM-2)$

Variable	Valor GVIF
V16_Amortizacion	1,397315
V38_AntiguedadCliente	1,084689
V46_Edad	1,033459
V60_NumeroInst_Adeuda	1,784055
V90_VariacionDeudaMicrocredito_U6M	1,073158
V91_VariacionDeudaTarjeta_U6M	1,084702
V17_CarteraRiesgo_Q(0; 0,10607]	1,528606
V17_CarteraRiesgo_Q(0,10607; Inf]	2,005504
V18_SaldoMMora_Mont(0,0034556; 0,042912]	1,299765
V18_SaldoMMora_Mont(0,042912; 0,059177]	1,680544
V18_SaldoMMora_Mont(0,059177, Inf]	2,861429
V29_PromAtraso_DC(0; 1]	2,078307
V29_PromAtraso_DC(1; 2]	2,626468
V29_PromAtraso_DC(2; 4]	3,776808
V29_PromAtraso_DC(4; 6]	3,053831
V29_PromAtraso_DC(6; Inf]	4,482136
V79_MoraPonderada_comp(0,42857; 3,05]	1,088482
V79_MoraPonderada_comp(3,05; Inf]	1,522092
V82_Saldo_MB_Comp(0,0052869; Inf]	1,378930

Fuente: Elaboración propia.

De la Tabla 3.11, se puede concluir que no existe multicolinealidad entre las variables explicativas paramétricas del modelo, ya que los valores del $GVIF$ para cada una de las variables son menores a 10.

En los GAM, incluso si dos variables no son colineales, pueden tener **concurrencia**, es decir, cuando los suavizados de dos variables explicativas están relacionadas entre sí de forma no lineal.

La definición de concurrencia se encuentra por primera vez en Buja et al. (1989)

y es esencialmente la forma no lineal de colinealidad. [Wood, Simon N., 2017], afirma que si hay convergencia, uno puede sentirse bastante seguro de los resultados incluso en presencia de concurrencia. A continuación, se muestra la concurrencia del modelo $GAM - 2$:

Tabla 3.12: Concurrencia - Modelo $(GAM - 2)$

	para	s(V31_AtrasoMax_U6_AC)	s(V81_Amortizacion_comp)
worst	0,9779774	0,8448724	0,4500307
observed	0,9779774	0,8347673	0,1993575
estimate	0,9779774	0,5727678	0,0532885

Fuente: Elaboración propia.

En la Tabla 3.12 se debe mirar el peor de los casos (worst), con valores entre 0 y 1 (donde 0 indica que no hay concurrencia).

- **Medidas de discriminación:** Para medir si el ajuste y discriminación del modelo son adecuados, se calcula los estadísticos: AUC, KS y coeficiente de $GINI$ tanto para la muestra de modelamiento como para la muestra de validación. En la Tabla 3.13 se muestran los valores calculados :

Tabla 3.13: Medidas de discriminación - Modelo $(GAM - 2)$

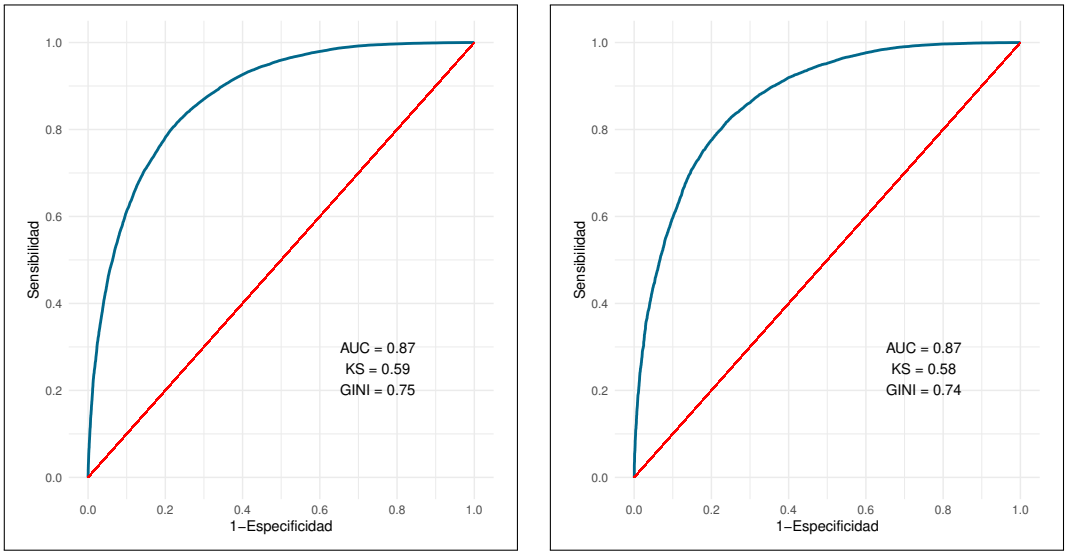
Estadístico	Muestra de Modelamiento	Muestra de Validación
AUC	0,8742493	0,8701103
KS	0,5855133	0,5783655
GINI	0,7484985	0,7402206

Fuente: Elaboración propia.

Podemos observar en la Tabla 3.13 que el estadístico AUC ($AUROC$) tanto en la muestra de modelamiento como de validación es mayor a $0{,}8$. Por lo que, podemos afirmar que el modelo propuesto presenta una correcta discriminación entre clientes Buenos y Malos. El valor del estadístico KS es de $0{,}58$ para la muestra de modelamiento y de $0{,}57$ para la muestra de validación, valores que [Anderson, 2007] considera adecuados. Motivo por el cual, el poder predictivo del modelo es alto. Por último, un coeficiente de $GINI$ mayor a $0{,}6$ para ambas muestras nos permite concluir que el rendimiento del modelo es más que satisfactorio.

En la Figura 3.9, se presenta la curva ROC junto con las medidas de discriminación para la muestra de modelamiento y validación.

Figura 3.9: Curva ROC muestra de modelamiento/validación - Modelo $(GAM - 2)$



Fuente: Elaboración propia.

- **Matriz de confusión:** Para construir la matriz de confusión, primero se establece el punto de corte óptimo del modelo. En una curva ROC, el punto de corte coincide con el punto en donde la distancia vertical entre la curva y la diagonal sea máxima.

  Empleando el punto de corte óptimo, cuyo valor es de $0,78$ y la muestra de modelamiento / validación se construyen las siguientes matrices de confusión junto con las métricas descritas a detalle en la sección (2.6.5), ver Tabla 3.14:

Tabla 3.14: Matriz de confusión y métricas - Modelo $(GAM - 2)$

Muestra de Modelamiento		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	16811	12529
1: Bueno	4535	49161

Métricas	Valor
Precisión	0,7945
Sensibilidad	0,7969
Especifidad	0,7875
Valor de predicción positivo	0,9155
Valor de predicción negativo	0,5730

Muestra de Validación		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	7120	5574
1: Bueno	1930	20964

Métricas	Valor
Precisión	0,7891
Sensibilidad	0,7900
Especifidad	0,7867
Valor de predicción positivo	0,9157
Valor de predicción negativo	0,5609

Fuente: Elaboración propia.

De la Tabla 3.14 se tiene que la Sensibilidad (% de Buenos clasificados correctamente) es de $79{,}69\,\%$ para la muestra de modelamiento y $79,00\,\%$ para la muestra de validación. La Especificidad (% de Malos clasificados correctamente) es de $78{,}75\,\%$ para la muestra de modelamiento y $78{,}67\,\%$ para la muestra de validación. Como podemos observar, los porcentajes de clasificación correcta son altos y muy similares en ambas muestras.

Además, el resto de métricas calculadas tienen también valores adecuados (en ambas muestras) según lo descrito en la sección (2.6.5). Las métricas calculadas ayudan a concluir que el modelo logístico aditivo generalizado posee un excelente poder de discriminación.

- **Tablas de desempeño:** En las tablas de desempeño (ver Tablas 3.15 y 3.16) para la muestra de modelamiento y validación observamos la distribución de clientes totales, clientes Buenos y clientes Malos por cada uno de los diez rangos de probabilidad de buen pagador.

  En ambos casos se puede observar que la tasa de clientes Buenos (TasaBuenos) en cada intervalo de probabilidad aumenta cuando la probabilidad aumenta y la tasa de clientes Malos (TasaMalos) disminuye cuando la probabilidad en cada intervalo

aumenta.

Tabla 3.15: Tabla de desempeño Muestra de modelamiento - Modelo $(GAM-2)$

Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Razón Buenos:Malos	
Buen Pagador	Num	Porc	PorcAcum	NumB	PorcB	PorcAcumB	NumM	PorcM	PorcAcumM	TasaBuenos	TasaMalos
[0,000; 0,227)	8306	10,00 %	10,00 %	803	1,3 %	1,3 %	7503	35,1 %	35,1 %	9,7 %	90,3 %
[0,227; 0,526)	8318	10,02 %	20,02 %	3338	5,4 %	6,7 %	4980	23,3 %	58,4 %	40,1 %	59,9 %
[0,526; 0,719)	8309	10,01 %	30,03 %	5204	8,4 %	15,1 %	3105	14,5 %	72,9 %	62,6 %	37,4 %
[0,719; 0,820)	8287	9,98 %	40,01 %	6267	10,2 %	25,3 %	2020	9,5 %	82,4 %	75,6 %	24,4 %
[0,820; 0,872)	8317	10,02 %	50,03 %	6938	11,2 %	36,5 %	1379	6,5 %	88,9 %	83,4 %	16,6 %
[0,872; 0,905)	8393	10,11 %	60,14 %	7460	12,1 %	48,6 %	933	4,4 %	93,3 %	88,9 %	11,1 %
[0,905; 0,930)	8085	9,74 %	69,88 %	7482	12,1 %	60,7 %	603	2,8 %	96,1 %	92,5 %	7,5 %
[0,930; 0,952)	8251	9,94 %	79,82 %	7831	12,7 %	73,4 %	420	2,0 %	98,1 %	94,9 %	5,1 %
[0,952; 0,973)	8609	10,37 %	90,19 %	8341	13,5 %	86,9 %	268	1,3 %	99,4 %	96,9 %	3,1 %
[0,973; 0,998]	8161	9,83 %	100,02 %	8026	13,0 %	99,9 %	135	0,6 %	100,0 %	98,3 %	1,7 %
Total	**83036**			**61690**			**21346**				

Fuente: Elaboración propia.

Tabla 3.16: Tabla de desempeño Muestra de validación - Modelo $(GAM-2)$

Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Tasa Buenos:Malos	
Buen Pagador	Num	Porc	PorcAcum	NumB	PorcB	PorcAcumB	NumM	PorcM	PorcAcumM	TasaBuenos	TasaMalos
[0,004; 0,231)	3560	10,00 %	10,00 %	404	1,5 %	1,5 %	3156	34,9 %	34,9 %	11,3 %	88,7 %
[0,231; 0,532)	3567	10,02 %	20,02 %	1513	5,7 %	7,2 %	2054	22,7 %	57,6 %	42,4 %	57,6 %
[0,532; 0,714)	3560	10,00 %	30,02 %	2207	8,3 %	15,5 %	1353	15,0 %	72,6 %	62,0 %	38,0 %
[0,714; 0,816)	3531	9,92 %	39,94 %	2631	9,9 %	25,4 %	900	9,9 %	82,5 %	74,5 %	25,5 %
[0,816; 0,870)	3607	10,14 %	50,08 %	3054	11,5 %	36,9 %	553	6,1 %	88,6 %	84,7 %	15,3 %
[0,870; 0,903)	3551	9,98 %	60,06 %	3151	11,9 %	48,8 %	400	4,4 %	93,0 %	88,7 %	11,3 %
[0,903; 0,929)	3548	9,97 %	70,03 %	3262	12,3 %	61,1 %	286	3,2 %	96,2 %	91,9 %	8,1 %
[0,929; 0,952)	3605	10,13 %	80,16 %	3422	12,9 %	74,0 %	183	2,0 %	98,2 %	94,9 %	5,1 %
[0,952; 0,972)	3433	9,65 %	89,81 %	3322	12,5 %	86,5 %	111	1,2 %	99,4 %	96,8 %	3,2 %
[0,972; 0,997]	3626	10,19 %	100,00 %	3572	13,5 %	100,0 %	54	0,6 %	100,0 %	98,5 %	1,5 %
Total	**35588**			**26538**			**9050**				

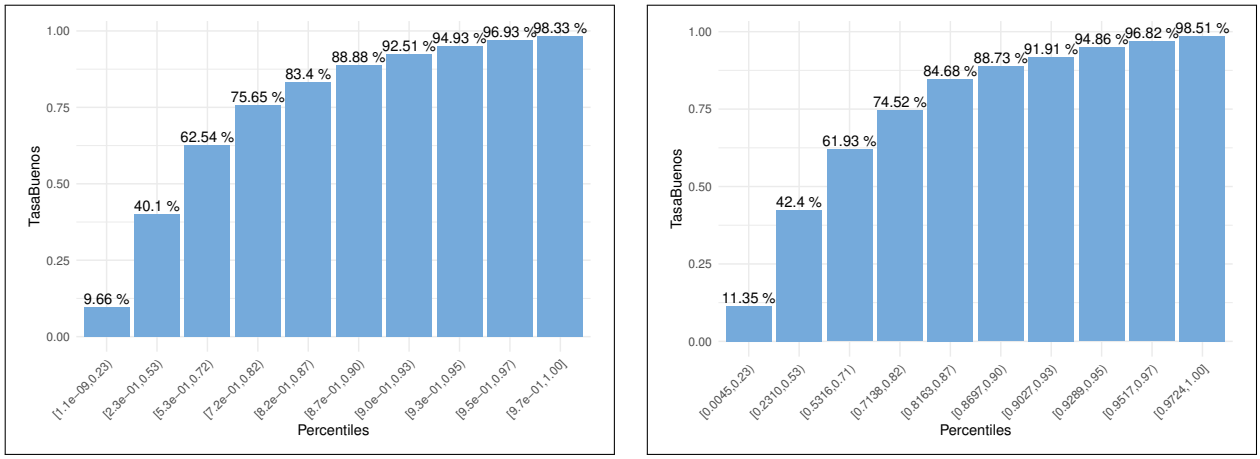
Fuente: Elaboración propia.

En las tablas anteriores se cumple con la condición descrita en la sección (2.6.6) de que en el decil más alto se concentre un porcentaje significativo de clientes Buenos, $98{,}3\,\%$ y apenas un $1{,}7\,\%$ de clientes Malos para la muestra de modelamiento y un $98{,}5\,\%$ de clientes Buenos; y $1,5\,\%$ de clientes Malos para la muestra de validación.

El modelo no presenta sobreajuste ya que la distribución de clientes Buenos y Malos no varía significativamente en ambas muestras. En la Figura 3.10 se grafica la Tasa

de Buenos por cada uno de los diez intervalos para la muestra de modelamiento y validación.

Figura 3.10: Tasa de Buenos por deciles muestra de modelamiento/validación - Modelo $(GAM - 2)$

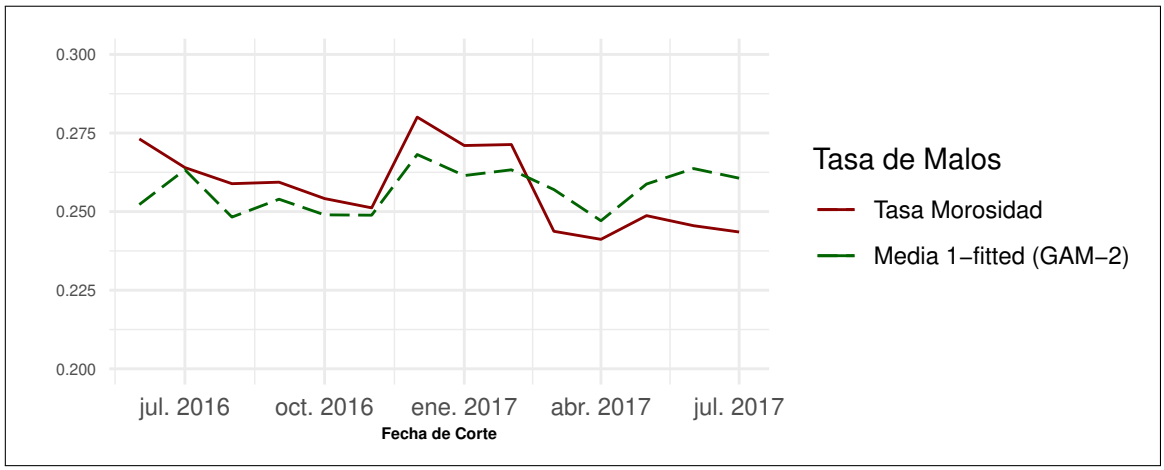


Fuente: Elaboración propia.

En donde se puede apreciar que el modelo logra segmentar muy bien a los clientes, con lo cual podemos concluir que nuestro modelo presenta un alto rendimiento de discriminación.

Por último en la Figura 3.11 se muestra la tasa de morosidad observada (Tasa Morosidad) y la tasa de morosidad pronosticada (Media 1-fitted) a lo largo del Punto de Observación (Fecha de Corte).

Figura 3.11: Tasa de morosidad a lo largo de la Fecha de Corte - Modelo $(GAM - 2)$



Fuente: Elaboración propia.

Del gráfico anterior, podemos observar el resultado de modelar la probabilidad de

incumplimiento (Morosidad). Se tiene como resultado que el modelo aditivo generalizado logra estimar exitosamente la tasa de morosidad observada.

Capítulo 4

Comparación entre modelo logístico aditivo generalizado y modelo de regresión logística

En el presente capítulo se construyen dos modelos empleando la técnica de regresión logística múltiple bajo procedimientos diferentes. Estos modelos se construyen para poder compararlos con el modelo logístico aditivo generalizado óptimo ($GAM-2$) con el fin de identificar la técnica y el modelo junto con la metodología empleada que mayor rendimiento posea tanto en predicción como en interpretación.

Para la construcción de ambos modelos se utilizan las mismas muestras de modelamiento y validación, esto hace que la comparación sea válida. Teniendo en cuenta estas consideraciones, procedemos a explicar el proceso de construcción de los modelos.

4.1 Construcción de los modelos de regresión logística

En esta sección se explica a detalle cómo se construye cada modelo. El primer modelo de regresión logística $(RL-M1)$ se construye con todas las variables que ingresaron al modelo logístico aditivo generalizado óptimo $(GAM-2)$. Se busca construir un modelo que posea un alto rendimiento y poder de discriminación como el modelo logístico aditivo generalizado; y que, además, sea interpretativo. Para lo cual, a

aquellas variables no paramétricas (Variables ajustadas mediante funciones suaves, $V31\_AtrasoMax\_U6\_AC$ y $V81\_Amortizacion\_comp$) se las categoriza o transforma observando la tendencia o comportamiento que posean en la Figura 3.4 y Figura 3.5 según corresponda.

Aplicando lo anteriormente mencionado, a las variables $V31\_AtrasoMax\_U6\_AC$ y $V81\_Amortizacion\_comp$ se les realizó una transformación cuadrátrica. En la Tabla 4.1 se muestra el modelo de regresión logística resultante, construido bajo el primer método $(RL - M1)$:

Tabla 4.1: Modelo de regresión logística $(RL - M1)$

Variable	Estimate	Std, Error z	value Pr(>|z|)
(Intercept)	4,789e+00	6,914e-02 69,270	<2e-16 ***
V16_Amortizacion	-3,592e+00	6,737e-02 -53,326	<2e-16 ***
V31_AtrasoMax_U6_AC	-7,417e-02	2,606e-03 -28,462	<2e-16 ***
I(V31_AtrasoMax_U6_AC^2)	4,008e-04	2,062e-05 19,435	<2e-16 ***
V38_AntiguedadCliente	2,432e-02	1,480e-03 16,434	<2e-16 ***
V46_Edad	7,391e-03	9,260e-04 7,982	1,44e-15 ***
V60_NumeroInst_Adeuda	-1,734e-01	1,084e-02 -15,992	<2e-16 ***
V81_Amortizacion_comp	5,730e-03	3,483e-03 1,645	0,099976 ,
I(V81_Amortizacion_comp^2)	-1,304e-04	3,525e-05 -3,698	0,000217 ***
V90_VariacionDeudaMicrocredito_U6M	-3,292e-01	3,199e-02 -10,292	<2e-16 ***
V91_VariacionDeudaTarjeta_U6M	-1,641e-01	3,089e-02 -5,312	1,09e-07 ***
V17_CarteraRiesgo_Q(0,0.10607]	-2,016e+00	5,134e-02 -39,264	<2e-16 ***
V17_CarteraRiesgo_Q(0.10607, Inf]	-2,522e+00	6,025e-02 -41,860	<2e-16 ***
V18_SaldoMMora_Mont(0.0034556,0.042912]	4,058e-01	4,746e-02 8,551	<2e-16 ***
V18_SaldoMMora_Mont(0.042912,0.059177]	3,661e-01	4,545e-02 8,055	7,97e-16 ***
V18_SaldoMMora_Mont(0.059177, Inf]	8,572e-01	5,428e-02 15,791	<2e-16 ***
V29_PromAtraso_DC(0,1]	-6,699e-01	3,073e-02 -21,803	<2e-16 ***
V29_PromAtraso_DC(1,2]	-9,465e-01	3,793e-02 -24,955	<2e-16 ***
V29_PromAtraso_DC(2,4]	-1,167e+00	4,163e-02 -28,026	<2e-16 ***
V29_PromAtraso_DC(4,6]	-1,434e+00	5,554e-02 -25,828	<2e-16 ***
V29_PromAtraso_DC(6, Inf]	-1,535e+00	6,838e-02 -22,451	<2e-16 ***
V79_MoraPonderada_comp(0.42857,3.05]	-2,274e-01	3,439e-02 -6,613	3,78e-11 ***
V79_MoraPonderada_comp(3.05, Inf]	-7,908e-01	2,744e-02 -28,821	<2e-16 ***
V82_Saldo_MB_Comp(0.0052869, Inf]	-7,736e-01	3,980e-02 -19,440	<2e-16 ***
Null deviance: 94656 on 83035 degrees of freedom			
Residual deviance: 61585 on 83012 degrees of freedom			
AIC: 61633			

Fuente: Elaboración propia.

La transformación realizada a las variables no paramétricas hizo que el modelo pre-

sentado en la Tabla 4.1 alcance un rendimiento y poder de discriminación más que satisfactorios. Por otro lado, podemos observar que la variable $V31\_AtrasoMax\_U6\_AC$ no es estadísticamente significativa con un nivel de confianza del $95\,\%$. Sin embargo, esta variable la podemos conservar al ser considerada como importante y porque todos los signos de los coeficientes son consistentes.

Para la contrucción del segundo modelo $(RL - M2)$ se incluyen inicialmente todas aquellas variables explicativas que superan el análisis exploratorio y los estadísticos Kolmogórov Smirnov y Valor de Información para finalmente encontrar un modelo óptimo mediante el algoritmo *backward* (ver sección 3.6). A continuación, se presenta el modelo de regresión logística resultante bajo el segundo método (método tradicional).

Tabla 4.2: Modelo de regresión logística $(RL - M2)$

Variable	Estimate	Std,	Error z	value Pr(>|z|)
(Intercept)	2,015679	0,210931	9,556	<2e-16 ***
V16_Amortizacion	-3,539492	0,067430	-52,491	<2e-16 ***
V31_AtrasoMax_U6_AC	-0,030749	0,001577	-19,501	<2e-16 ***
V38_AntiguedadCliente	0,024335	0,001478	16,460	<2e-16 ***
V46_Edad	0,007521	0,000922	8,157	3,44e-16 ***
V60_NumeroInst_Adeuda	-0,163860	0,010589	-15,475	<2e-16 ***
V68_TiempoHistorialCrediticioSF_12	0,053023	0,016650	3,184	0,00145 **
V81_Amortizacion_comp	-0,007312	0,002326	-3,144	0,00167 **
V90_VariacionDeudaMicrocredito_U6M	-0,320221	0,031818	-10,064	<2e-16 ***
V91_VariacionDeudaTarjeta_U6M	-0,172607	0,030731	-5,617	1,95e-08 ***
V18_SaldoMMora_Mont(0.0034556,0.042912]	0,357181	0,046550	7,673	1,68e-14 ***
V18_SaldoMMora_Mont(0.042912,0.059177]	0,310256	0,044124	7,031	2,04e-12 ***
V18_SaldoMMora_Mont(0.059177, Inf]	0,659015	0,050126	13,147	<2e-16 ***
V25_Porc_CuotasPag(0.88889,0.95652]	-0,337938	0,059885	-5,643	1,67e-08 ***
V25_Porc_CuotasPag(0.95652, Inf]	2,050268	0,051135	40,095	<2e-16 ***
V29_PromAtraso_DC(0,1]	-0,820221	0,029792	-27,532	<2e-16 ***
V29_PromAtraso_DC(1,2]	-1,216826	0,035585	-34,195	<2e-16 ***
V29_PromAtraso_DC(2,4]	-1,566700	0,037147	-42,176	<2e-16 ***
V29_PromAtraso_DC(4,6]	-1,972121	0,049647	-39,723	<2e-16 ***
V29_PromAtraso_DC(6, Inf]	-2,219341	0,061883	-35,864	<2e-16 ***
V79_MoraPonderada_comp(0.42857,3.05]	-0,242604	0,034286	-7,076	1,48e-12 ***
V79_MoraPonderada_comp(3.05, Inf]	-0,812966	0,027317	-29,760	<2e-16 ***
V82_Saldo_MB_Comp(0.0052869, Inf]	-0,773115	0,039554	-19,546	<2e-16 ***

Null deviance: 94656 on 83035 degrees of freedom
Residual deviance: 62075 on 83013 degrees of freedom
AIC: 62121

Fuente: Elaboración propia.

Podemos observar en la Tabla 4.2, que todas las variables del modelo son estadísticamente significativas y considerando el significado de cada variable (ANEXO 1) tenemos que los signos de los coeficientes son correctos.

4.2 Resultados y evaluación de los modelos de regresión logística

En la presente sección se analiza la calidad de discriminación y predicción de los dos modelos de regresión logística construidos mediante los métodos explicados anteriormente. Para evaluar el rendimiento tanto del modelo logístico aditivo generalizado como de los modelos de regresión logística construidos se emplea la muestra de validación y las técnicas estadísticas vistas en la sección (2.6).

Comenzamos presentando los resultados obtenidos de cada técnica estadística para el modelo de regresión logística construido bajo el primer método $(RL - M1)$.

- **Multicolinealidad:** Se realizó el cálculo del factor de inflación de varianza generalizada $(GVIF)$ para cada variable del modelo y se obtuvieron los siguientes valores:

Tabla 4.3: $GVIF$ - Modelo $(RL - M1)$

Variable	Valor GVIF
V16_Amortizacion	1,371907
V31_AtrasoMax_U6_AC	9,546295
I(V31_AtrasoMax_U6_AC^2)	4,471851
V38_AntiguedadCliente	1,079672
V46_Edad	1,029900
V60_NumeroInst_Adeuda	1,322714
V81_Amortizacion_comp	1,802290
I(V81_Amortizacion_comp^2)	1,616219
V90_VariacionDeudaMicrocredito_U6M	1,042482
V91_VariacionDeudaTarjeta_U6M	1,085339
V17_CarteraRiesgo_Q	3,021399
V18_SaldoMMora_Mont	4,462563
V29_PromAtraso_DC	4,221127
V79_MoraPonderada_comp	1,475289
V82_Saldo_MB_Comp	1,378121

Fuente: Elaboración propia.

Con los resultados mostrados en la Tabla 4.3, se verifica que no existe multicolinealidad entre las variables explicativas del modelo $(RL - M1)$. Podemos observar que dos variables poseen un valor $GVIF$ inflado, esto se debe a que estas variables forman parte del modelo con una transformación cuadrática; por lo que esto no representa ningún problema ya que todos los valores son menores que 10.

- **Medidas de discriminación:** Para validar la calidad de discriminación y predicción del modelo , en la Tabla 4.4 se muestran los estadísticos calculados. En donde podemos observar que los valores AUC, KS y $GINI$ son adecuados según lo expuesto en la sección (2.6), con lo cual podemos concluir que el modelo presenta adecuadas características de discriminación y predicción.

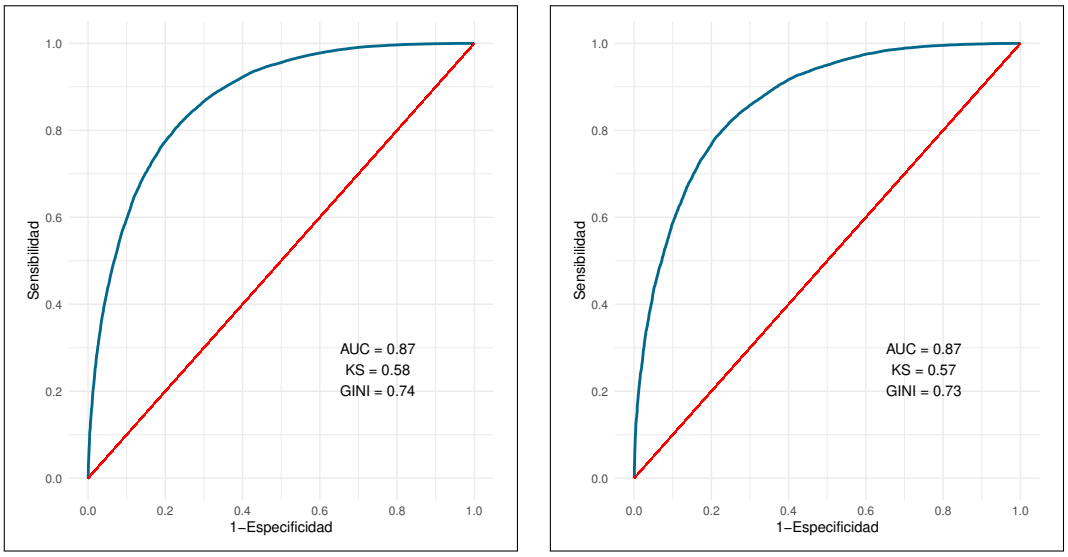
Tabla 4.4: Medidas de discriminación - Modelo $(RL - M1)$

Estadístico	Muestra de Modelamiento	Muestra de Validación
AUC	0,8703792	0,8658855
KS	0,57824	0,572615
GINI	0,7407585	0,7317711

Fuente: Elaboración propia.

En la Figura 4.1, se presenta el gráfico de la curva ROC junto con las medidas de discriminación para la muestra de modelamiento y validación.

Figura 4.1: Curva ROC muestra de modelamiento/validación - Modelo $(RL - M1)$



Fuente: Elaboración propia.

- **Matriz de confusión:** En la Tabla 4.5 se presenta la matriz de confusión con el valor del punto de corte igual a $0{,}78$, punto de corte óptimo con el cual se tiene que el $79,26\,\%$ y $78,29\,\%$ de los clientes Buenos y Malos respectivamente fueron clasificados correctamente en la muestra de modelamiento y valores cercanos en la muestra de validación.

Tabla 4.5: Matriz de confusión y métricas - Modelo $(RL - M1)$

Muestra de Modelamiento		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	16711	12793
1: Bueno	4635	48897

Métricas	Valor
Precisión	0,7901
Sensibilidad	0,7926
Especifidad	0,7829
Valor de predicción positivo	0,9134
Valor de predicción negativo	0,5664

Muestra de Validación		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	7097	5640
1: Bueno	1953	20898

Métricas	Valor
Precisión	0,7866
Sensibilidad	0,7875
Especifidad	0,7842
Valor de predicción positivo	0,9145
Valor de predicción negativo	0,5572

Fuente: Elaboración propia.

- **Tablas de desempeño:** En las tablas de desempeño (Ver Tablas 4.6 y 4.7) para la muestra de modelamiento y validación observamos que la tasa de clientes Buenos (TasaBuenos) en cada intervalo de probabilidad aumenta cuando la probabilidad aumenta con un crecimiento estricto, sucede lo contrario con la Tasa de Malos (TasaMalos).

Tabla 4.6: Tabla de desempeño muestra de modelamiento - Modelo $(RL - M1)$

Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Razón Buenos:Malos	
Buen Pagador	**Num**	**Porc**	**PorcAcum**	**NumB**	**PorcB**	**PorcAcumB**	**NumM**	**PorcM**	**PorcAcumM**	**TasaBuenos**	**TasaMalos**
[0,000; 0,231)	8294	9,99 %	9,99 %	895	1,5 %	1,5 %	7399	34,7 %	34,7 %	10,8 %	89,2 %
[0,231; 0,539)	8302	10,00 %	19,99 %	3310	5,4 %	6,9 %	4992	23,4 %	58,1 %	39,9 %	60,1 %
[0,539; 0,718)	8286	9,98 %	29,97 %	5204	8,4 %	15,3 %	3082	14,4 %	72,5 %	62,8 %	37,2 %
[0,718; 0,818)	8346	10,05 %	40,02 %	6281	10,2 %	25,5 %	2065	9,7 %	82,2 %	75,3 %	24,7 %
[0,818; 0,869)	8303	10,00 %	50,02 %	6923	11,2 %	36,7 %	1380	6,5 %	88,7 %	83,4 %	16,6 %
[0,869; 0,901)	8366	10,08 %	60,10 %	7470	12,1 %	48,8 %	896	4,2 %	92,9 %	89,3 %	10,7 %
[0,901; 0,926)	8073	9,72 %	69,82 %	7410	12,0 %	60,8 %	663	3,1 %	96,0 %	91,8 %	8,2 %
[0,926; 0,949)	8457	10,18 %	80,00 %	8014	13,0 %	73,8 %	443	2,1 %	98,1 %	94,8 %	5,2 %
[0,949; 0,970)	8286	9,98 %	89,98 %	8004	13,0 %	86,8 %	282	1,3 %	99,4 %	96,6 %	3,4 %
[0,970; 1,000]	8323	10,02 %	100,00 %	8179	13,3 %	100,1 %	144	0,7 %	100,1 %	98,3 %	1,7 %
Total	**83036**			**61690**			**21346**				

Fuente: Elaboración propia.

Tabla 4.7: Tabla de desempeño muestra de validación - Modelo $(RL - M1)$

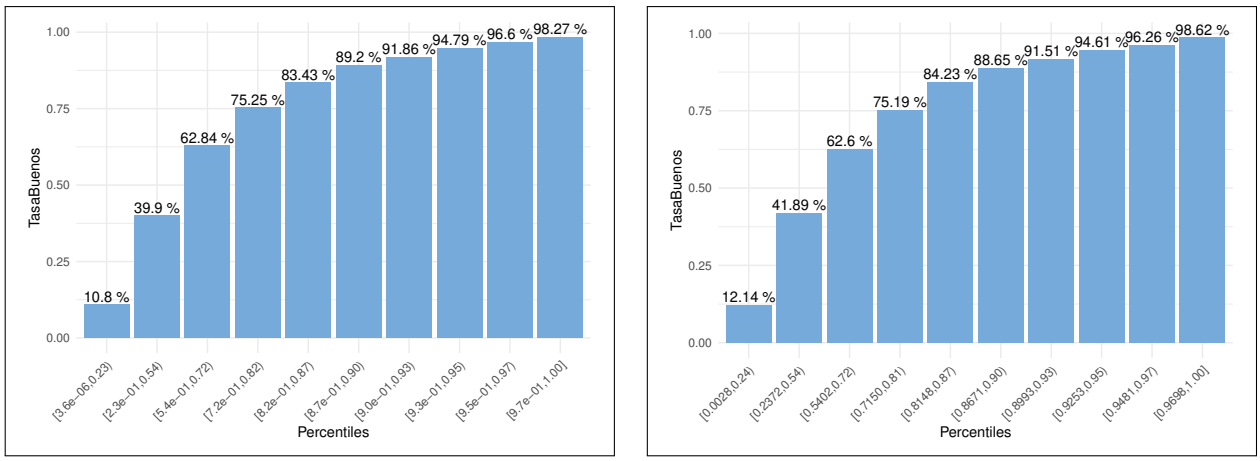
Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Tasa Buenos:Malos	
Buen Pagador	Num	Porc	PorcAcum	NumB	PorcB	PorcAcumB	NumM	PorcM	PorcAcumM	TasaBuenos	TasaMalos
[0,003; 0.237)	3559	10,0 %	10,0 %	432	1,6 %	1,6 %	3127	34,6 %	34,6 %	12,1 %	87,9 %
[0,237; 0,540)	3559	10,0 %	20,0 %	1491	5,6 %	7,2 %	2068	22,9 %	57,5 %	41,9 %	58,1 %
[0,540; 0,715)	3559	10,0 %	30,0 %	2228	8,4 %	15,6 %	1331	14,7 %	72,2 %	62,6 %	37,4 %
[0,715; 0,815)	3559	10,0 %	40,0 %	2676	10,1 %	25,7 %	883	9,8 %	82,0 %	75,2 %	24,8 %
[0,815; 0,867)	3558	10,0 %	50,0 %	2997	11,3 %	37,0 %	561	6,2 %	88,2 %	84,2 %	15,8 %
[0,867; 0,899)	3559	10,0 %	60,0 %	3155	11,9 %	48,9 %	404	4,5 %	92,7 %	88,6 %	11,4 %
[0,899; 0,925)	3559	10,0 %	70,0 %	3257	12,3 %	61,2 %	302	3,3 %	96,0 %	91,5 %	8,5 %
[0,925; 0,948)	3559	10,0 %	80,0 %	3367	12,7 %	73,9 %	192	2,1 %	98,1 %	94,6 %	5,4 %
[0,948; 0,970)	3559	10,0 %	90,0 %	3426	12,9 %	86,8 %	133	1,5 %	99,6 %	96,3 %	3,7 %
[0,970; 0,996]	3558	10,0 %	100,0 %	3509	13,2 %	100,0 %	49	0,5 %	100,1 %	98,6 %	1,4 %
Total	**35588**			**26538**			**9050**				

Fuente: Elaboración propia.

La distribución de clientes Buenos y Malos no varía significativamente en ambas muestras, lo que indica que el modelo no presenta sobreajuste.

En el siguiente gráfico se presenta la Tasa de Buenos para los diez intervalos de probabilidad para la muestra de modelamiento y validación.

Figura 4.2: Tasa de Buenos por deciles muestra de modelamiento/validación - Modelo $(RL - M2)$

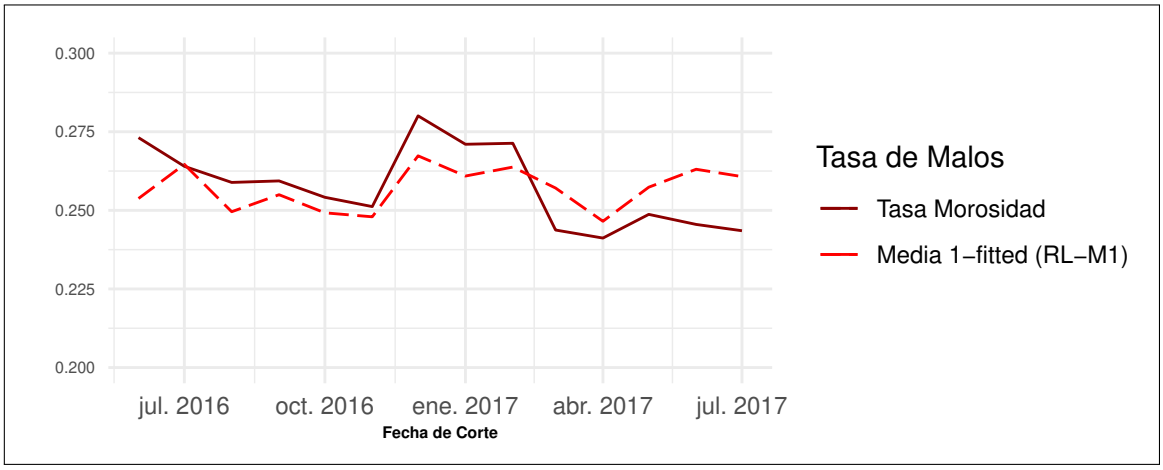


Fuente: Elaboración propia.

En la Figura 4.2 se observa que el modelo segmenta a los clientes adecuadamente, de esto se puede concluir que el modelo posee un alto rendimiento de discriminación.

A continuación, se muestra el gráfico de la tasa de morosidad observada (Tasa Morosidad) y la tasa de morosidad pronosticada (Media 1-fitted) a lo largo del Punto de Observación (Fecha de Corte).

Figura 4.3: Tasa de morosidad a lo largo de la Fecha de Corte - Modelo $(RL - M1)$



Fuente: Elaboración propia.

En la Figura 4.3, si nos fijamos en la probabilidad de incumplimiento (tasa de morosidad pronosticada), podemos observar que el Modelo $(RL - M1)$ pronostica con precisión la tasa de morosidad observada.

A continuación, se presentan los resultados de cada técnica estadística obtenidos para evaluar el rendimiento del modelo de regresión logística construido bajo el segundo método $(RL - M2)$, **método tradicional**.

- **Multicolinealidad:** Se obtuvieron los siguientes valores del factor de inflación de varianza generalizada $(GVIF)$ para el Modelo $(RL - M2)$:

Tabla 4.8: $GVIF$ - Modelo $(RL - M2)$

Variable	Valor GVIF
V16_Amortizacion	1,393234
V31_AtrasoMax_U6_AC	3,102116
V38_AntiguedadCliente	1,085838
V46_Edad	1,029915
V60_NumeroInst_Adeuda	1,274204
V68_TiempoHistorialCrediticioSF_12	1,053737
V81_Amortizacion_comp	1,087260
V90_VariacionDeudaMicrocredito_U6M	1,041893
V91_VariacionDeudaTarjeta_U6M	1,086216
V18_SaldoMMora_Mont	3,252459
V25_Porc_CuotasPag	2,190548
V29_PromAtraso_DC	3,183952
V79_MoraPonderada_comp	1,478144
V82_Saldo_MB_Comp	1,378902

Fuente: Elaboración propia.

Los resultados de la Tabla 4.8, muestran que no existe multicolinealidad entre las variables explicativas del modelo $(RL - M2)$, pues los valores del $GVIF$ para cada una de las variables son bajos menores de 5.

- **Medidas de discriminación:** El modelo $(RL - M2)$ estimado presenta un AUC de $0{,}8684$, un KS de $0{,}5748$ y un $GINI$ de $0{,}7367$ (ver Tabla 4.9) y valores muy similares en la muestra de validación, lo que permite concluir que el modelo tiene un buen ajuste, poder de discriminación y predicción.

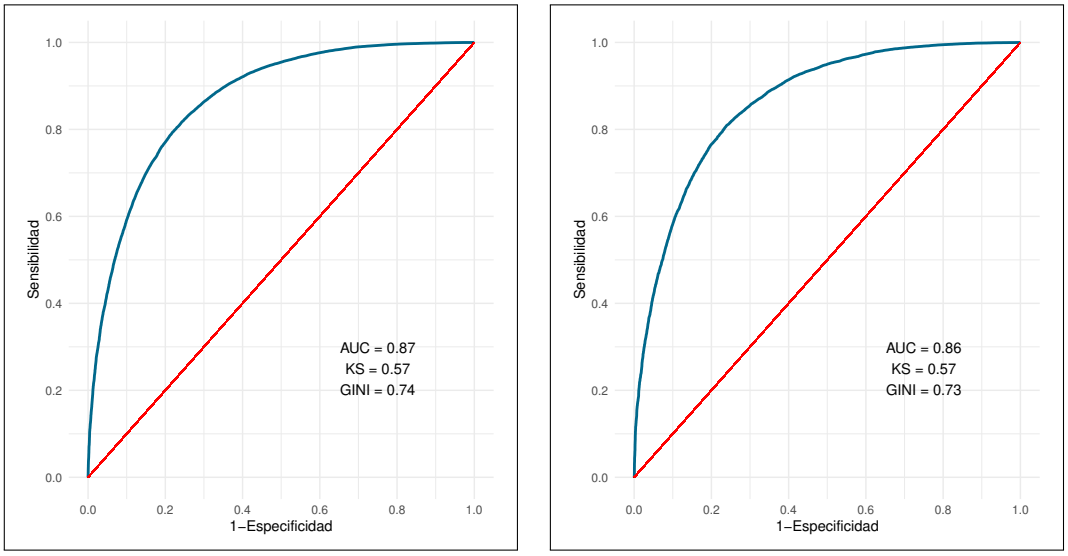
Tabla 4.9: Medidas de discriminación - Modelo $(RL - M2)$

Estadístico	Muestra de Modelamiento	Muestra de Validación
AUC	0,868366	0,8637951
KS	0,5748226	0,5697369
GINI	0,7367321	0,7275902

Fuente: Elaboración propia.

Se presenta el gráfico de la curva ROC junto con las medidas de discriminación para la muestra de modelamiento y validación.

Figura 4.4: Curva ROC muestra de modelamiento/validación - Modelo $(RL - M2)$



Fuente: Elaboración propia.

- **Matriz de confusión:** La matriz de confusión que se presenta en la Tabla 4.10 con un punto de corte igual a $0{,}78$ muestra que un $79,03\,\%$ y un $78,38\,\%$ de los clientes Buenos y Malos respectivamente fueron clasificados correctamente, estos valores son muy similares a los obtenidos para la muestra de validación, lo cual nos permite concluir que el modelo presenta un aceptable poder de clasificación.

Tabla 4.10: Matriz de confusión y métricas - Modelo $(RL - M2)$

Muestra de Modelamiento		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	16730	12934
1: Bueno	4616	48756

Métricas	**Valor**
Precisión	0,7886
Sensibilidad	0,7903
Especificidad	0,7838
Valor de predicción positivo	0,9135
Valor de predicción negativo	0,5640

Muestra de Validación		
Punto de Corte: 0,78	**Real**	
Pronóstico	**0: Malo**	**1: Bueno**
0: Malo	7089	5754
1: Bueno	1961	20784

Métricas	**Valor**
Precisión	0,7832
Sensibilidad	0,7832
Especificidad	0,7833
Valor de predicción positivo	0,9138
Valor de predicción negativo	0,5520

Fuente: Elaboración propia.

- **Tablas de desempeño:** Las tablas de desempeño del modelo $(RL - M2)$ (Tabla 4.6

y 4.7) para la muestra de modelamiento y validación muestran que el porcentaje de clientes Buenos (TasaBuenos) en cada rango de la probabilidad estimada aumenta con el aumento de la probabilidad con un crecimiento estricto, mientras que la tasa de clientes Malos (TasaMalos) disminuye con un decrecimiento estricto.

Tabla 4.11: Tabla de desempeño muestra de modelamiento - Modelo $(RL - M2)$

Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Razón Buenos:Malos	
Buen Pagador	Num	Porc	PorcAcum	NumB	PorcB	PorcAcumB	NumM	PorcM	PorcAcumM	TasaBuenos	TasaMalos
[0,000; 0,236)	8316	10,01 %	10,01 %	964	1,6 %	1,6 %	7352	34,4 %	34,4 %	11,6 %	88,4 %
[0,236; 0,543)	8292	9,99 %	20,00 %	3307	5,4 %	7,0 %	4985	23,4 %	57,8 %	39,9 %	60,1 %
[0,543; 0,717)	8281	9,97 %	29,97 %	5209	8,4 %	15,4 %	3072	14,4 %	72,2 %	62,9 %	37,1 %
[0,717; 0,816)	8310	10,01 %	39,98 %	6239	10,1 %	25,5 %	2071	9,7 %	81,9 %	75,1 %	24,9 %
[0,816; 0,867)	8242	9,93 %	49,91 %	6871	11,1 %	36,6 %	1371	6,4 %	88,3 %	83,4 %	16,6 %
[0,867; 0,899)	8298	9,99 %	59,90 %	7367	11,9 %	48,5 %	931	4,4 %	92,7 %	88,8 %	11,2 %
[0,899; 0,925)	8381	10,09 %	69,99 %	7714	12,5 %	61,0 %	667	3,1 %	95,8 %	92,0 %	8,0 %
[0,925; 0,948)	8397	10,11 %	80,10 %	7922	12,8 %	73,8 %	475	2,2 %	98,0 %	94,3 %	5,7 %
[0,948; 0,969)	8157	9,82 %	89,92 %	7881	12,8 %	86,6 %	276	1,3 %	99,3 %	96,6 %	3,4 %
[0,969; 0,997]	8362	10,07 %	99,99 %	8216	13,3 %	99,9 %	146	0,7 %	100,0 %	98,3 %	1,7 %
Total	**83036**			**61690**			**21346**				

Fuente: Elaboración propia.

Tabla 4.12: Tabla de desempeño muestra de validación - Modelo $(RL - M2)$

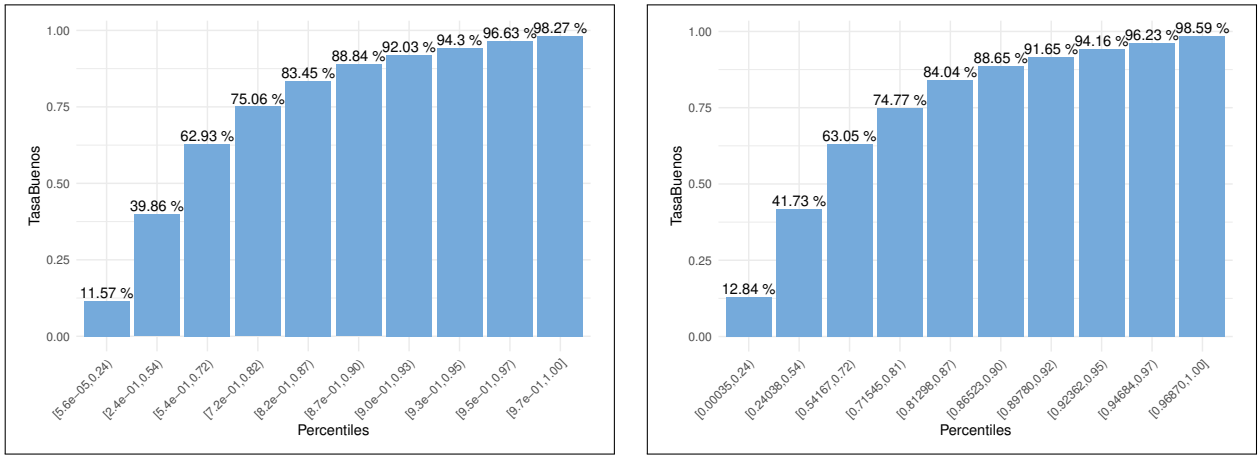
Probabilidad	Clientes Totales			Clientes Buenos			Clientes Malos			Tasa Buenos:Malos	
Buen Pagador	Num	Porc	PorcAcum	NumB	PorcB	PorcAcumB	NumM	PorcM	PorcAcumM	TasaBuenos	TasaMalos
[0,000; 0,240)	3559	10,0 %	10,0 %	457	1,7 %	1,7 %	3102	34,3 %	34,3 %	12,8 %	87,2 %
[0,240; 0,542)	3559	10,0 %	20,0 %	1485	5,6 %	7,3 %	2074	22,9 %	57,2 %	41,7 %	58,3 %
[0,542; 0,715)	3559	10,0 %	30,0 %	2244	8,5 %	15,8 %	1315	14,5 %	71,7 %	63,1 %	36,9 %
[0,715; 0,813)	3559	10,0 %	40,0 %	2661	10,0 %	25,8 %	898	9,9 %	81,6 %	74,8 %	25,2 %
[0,813; 0,865)	3558	10,0 %	50,0 %	2990	11,3 %	37,1 %	568	6,3 %	87,9 %	84,0 %	16,0 %
[0,865; 0,898)	3559	10,0 %	60,0 %	3155	11,9 %	49,0 %	404	4,5 %	92,4 %	88,6 %	11,4 %
[0,898; 0,924)	3559	10,0 %	70,0 %	3262	12,3 %	61,3 %	297	3,3 %	95,7 %	91,7 %	8,3 %
[0,924; 0,947)	3559	10,0 %	80,0 %	3351	12,6 %	73,9 %	208	2,3 %	98,0 %	94,2 %	5,8 %
[0,947; 0,969)	3559	10,0 %	90,0 %	3425	12,9 %	86,8 %	134	1,5 %	99,5 %	96,2 %	3,8 %
[0,969; 0,995]	3558	10,0 %	100,0 %	3508	13,2 %	100,0 %	50	0,6 %	100,1 %	98,6 %	1,4 %
Total	**35588**			**26538**			**9050**				

Fuente: Elaboración propia.

El modelo no presenta sobreajuste ya que la distribución de clientes Buenos y Malos no varía significativamente en ambas muestras.

En la Figura 4.5 se grafica la Tasa de Buenos para los diez intervalos de probabilidad empleando la muestra de modelamiento y validación.

Figura 4.5: Tasa de Buenos por deciles muestra de modelamiento/validación - Modelo $(RL - M2)$

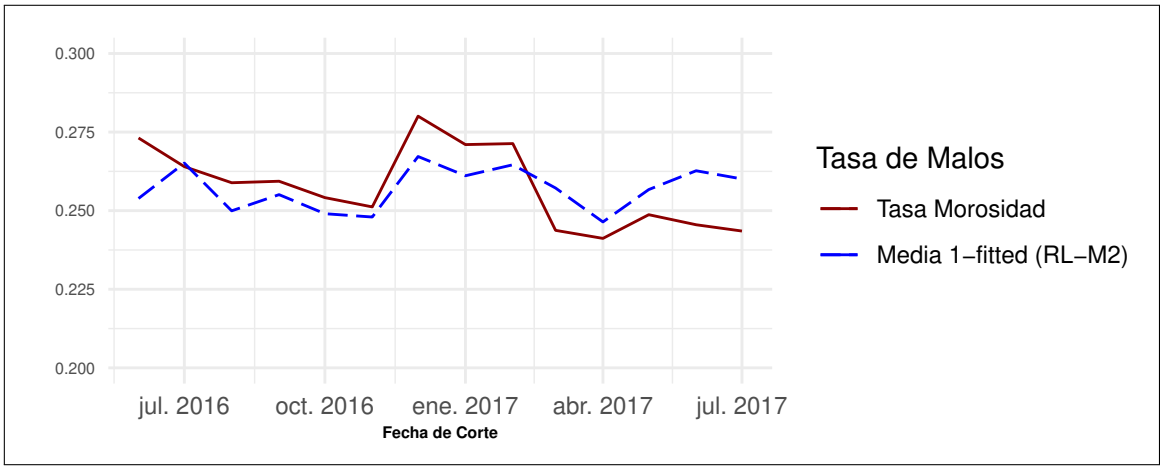


Fuente: Elaboración propia.

Del gráfico, con la muestra de modelamiento podemos observar que en el rango más alto de la probabilidad estimada el $98,27\,\%$ son clientes Buenos y apenas el $1,83\,\%$ son clientes Malos, mientras que en el rango más bajo únicamente el $11,57\,\%$ de los clientes son Buenos y el $88,47\,\%$ son clientes Malos. Esto nos permite concluir que el modelo posee buen rendimiento de discriminación.

Por último, se presenta el gráfico de la tasa de morosidad observada (Tasa Morosidad) y la tasa de morosidad pronosticada (Media 1-fitted) a lo largo del Punto de observación (Fecha de Corte).

Figura 4.6: Tasa de morosidad a lo largo de la Fecha de Corte - Modelo $(RL - M2)$



Fuente: Elaboración propia.

En la Figura 4.6 se observa que el Modelo $(RL - M2)$ logra pronosticar de menara

adecuada la tasa de morosidad observada.

4.3 Comparación Modelos Scoring Crediticio

En las secciones anteriores se construyeron los modelos y se presentaron las pruebas de validación de los mismos. Esto nos permitirá cumplir con el objetivo principal de este trabajo, establecer modelos estadísticos robustos mediante la técnica de Regresión Logística y el algoritmo Modelos Aditivos Generalizados y comparar los resultados obtenidos.

Para realizar esta comparación, se emplean las técnicas estadísticas detalladas en la sección (2.6).

4.3.1 Poder de discriminación

En la Tabla 4.13, se muestran el estadístico KS, el índice AUC y el coeficiente de $GINI$. Se observa que los indicadores calculados presentan un valor superior en el Modelo $(GAM-2)$. Además, los indicadores calculados para el Modelo $(RL-M1)$ son los que más se aproximan al Modelo $(GAM-2)$.

Tabla 4.13: Comparación medidas de discriminación

Estadístico	Modelo (GAM-2)	Modelo (RL-M1)	Modelo (RL-M2)
AUC	0,8742493	0,8703792	0,868366
KS	0,5855133	0,57824	0,5748226
GINI	0,7484985	0,7407585	0,7367321

Fuente: Elaboración propia.

4.3.2 Matriz de confusión e indicadores de eficiencia

Para comparar los modelos se construyen las matrices de confusión con el valor de los dos puntos de corte óptimos $(0,78)$. Comparamos el porcentaje de clasificación correcta comparando los valores calculados para los indicadores de eficiencia (precisión, sensibilidad, especificidad, valor de predicción positivo y negativo).

Tabla 4.14: Comparación matriz de confusión y métricas

Modelo (GAM)				Modelo (RL-M1)				Modelo (RL-M2)			
PC: 0,78		Real		PC: 0.78		Real		PC: 0.78		Real	
		0: Malo	1: Bueno			0: Malo	1: Bueno			0: Malo	1: Bueno
Pronóstico	0: Malo	16811	12529		0: Malo	16711	12793		0: Malo	16730	12934
	1: Bueno	4535	49161		1: Bueno	4635	48897		1: Bueno	4616	48756

Métricas	Valor	Valor	Valor
Precisión	0,7945	0,7901	0,7886
Sensibilidad	0,7969	0,7926	0,7903
Especifidad	0,7875	0,7829	0,7838
Valor de predicción positivo	0,9155	0,9134	0,9135
Valor de predicción negativo	0,5730	0,5664	0,5640

Fuente: Elaboración propia.

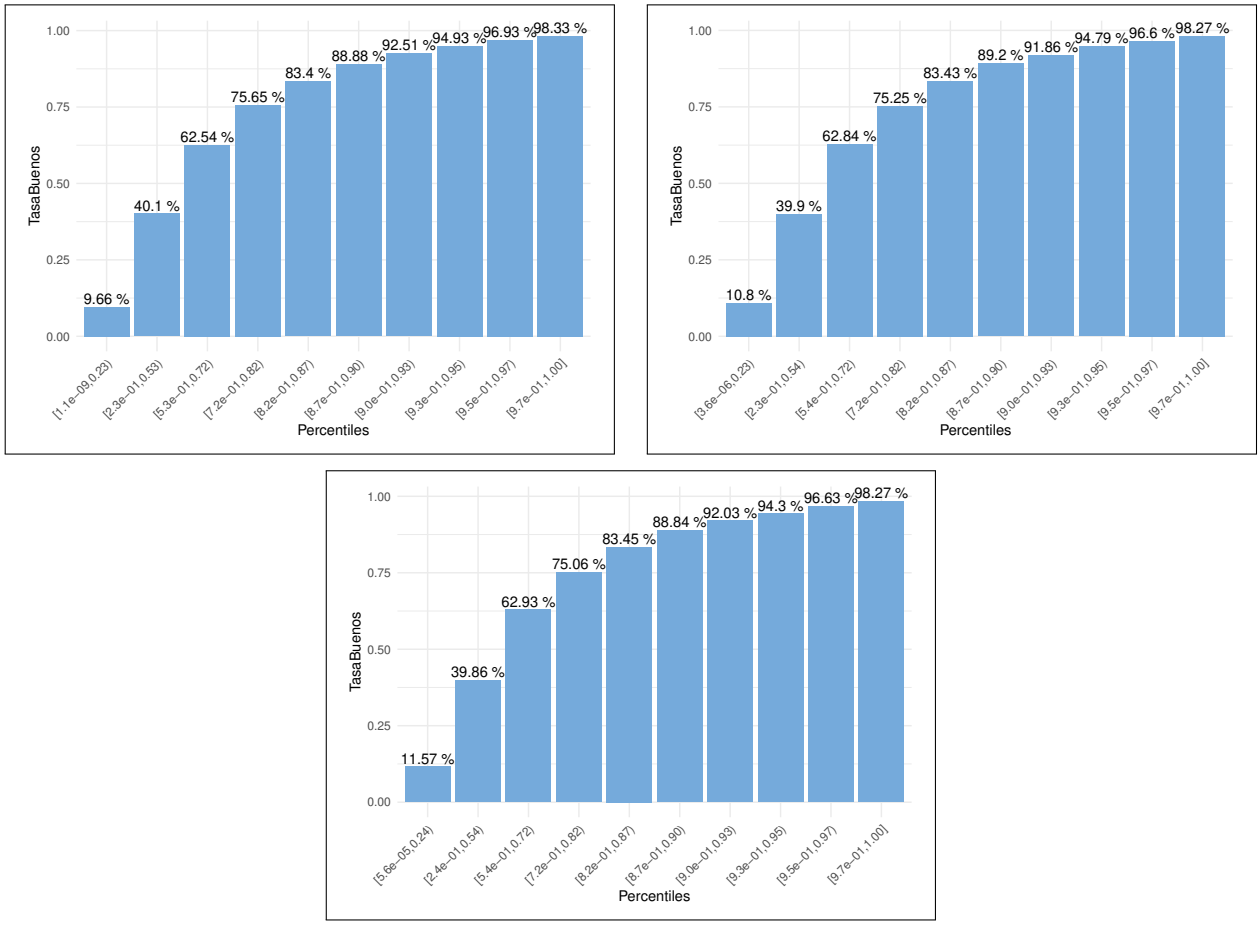
Analizando la Tabla 4.14 se puede observar que el Modelo $(GAM-2)$ tiene un $79,69\,\%$ y un $78,75\,\%$ de clientes Buenos y Malos clasificados correctamente, valores superiores a los dos modelos restantes. Sin embargo, el modelo logístico $(RL-M1)$ con un $79,26\,\%$ de clientes Buenos clasificados correctamente es el que más se aproxima al modelo logístico aditivo generalizado $(GAM-2)$.

4.3.3 Tablas de desempeño

La discriminación que logra un modelo se puede evaluar mediante una tabla de desempeño. Comparando el porcentaje de clientes Buenos (TasaBuenos) en cada rango de la probabilidad estimada no se observan mayores diferencias en los extremos de los diez intervalos. Sin embargo, en el gráfico de la Figura 4.7 se puede observar que el modelo logístico aditivo generalizado $(GAM-2)$ tiene un ordenamiento más parsimonioso que los dos modelos restantes. En otras palabras, este modelo segmenta de manera más adecuada a los clientes Buenos.

Por otro lado, el gráfico de deciles del modelo de regresión logística contruido bajo el primer método $(RL-M1)$ es el que más se aproxima (ver Figura 4.7).

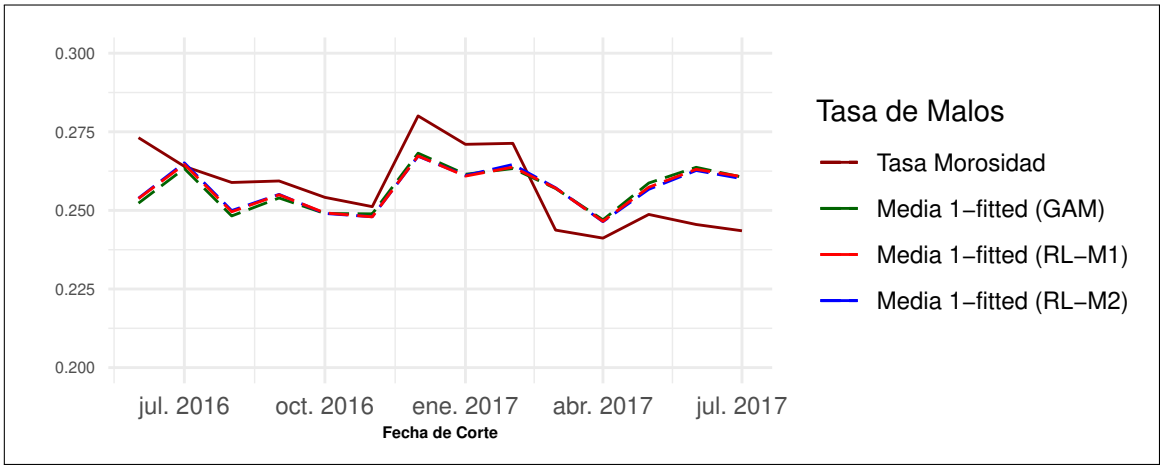
Figura 4.7: Tasa de Buenos por deciles - Modelos $(GAM-2)$, $(RL-M1)$ y $(RL-M2)$



Fuente: Elaboración propia.

Lo mencionado en el párrafo anterior se puede comprobar observando el gráfico de la tasa de morosidad observada (Tasa Morosidad) y la tasa de morosidad pronosticada (Media 1-fitted) a lo largo del Punto de Observación (Fecha de Corte) construido para los tres modelos.

Figura 4.8: Tasa de morosidad a lo largo de la Fecha de Corte - Modelos $(GAM-2)$, $(RL-M1)$ y $(RL-M2)$



Fuente: Elaboración propia.

Del gráfico anterior, se puede concluir que el modelo de regresión logística construido bajo el primer método $(RL-M1)$ pronostica de mejor manera que el modelo regresión logística construido bajo el método tradicional.

Por último, se concluye que los tres modelos cumplen un buen desempeño ya que la Tasa de Buenos tiene un ordenamiento creciente. Esto nos permite concluir que el poder de predicción y discriminación de los tres modelos presentados es alto.

Capítulo 5

Implementación de la Metodología Analítica en R

En el capítulo 3 (Metodología Analítica), se explicó cada uno de los pasos a seguir para la construcción de modelos estadísticos robustos. En este capítulo, con ayuda del software estadístico **R** se presenta el algoritmo implementado, el mismo que partiendo de una Base de datos permite ejecutar de manera automática cada uno de los pasos expuestos. Se empieza realizando una breve introducción al lenguaje de programación **R**.

5.1 Lenguaje de programación estadístico R

El software estadístico **R** es un lenguaje y un ambiente de programación libre, especialmente desarrollado para ser empelado en el análisis y tratamiento de datos, análisis estadístico y creación de gráficos de calidad. Esto es posible gracias a la amplia variedad de técnicas estadísticas que dispone, por ejemplo: tests estadísticos, modelos lineales y no lineales, algoritmos de clasificación y agrupamiento, análisis de series temporales, etc. Técnicas que pueden ampliarse fácilmente mediante paquetes, librerías o construyendo nuestras propias funciones.

R nació en 1993 en Auckland. Sin embargo, si se remonta a sus bases iniciales, puede decirse que inició con un lenguaje previo llamado S, creado por John Chambers y colaboradores en Bell Laboratories durante la década de 1970.

Por otro lado, además de ser un software libre, es de código abierto (Open Source) parte del proyecto GNU **GPL** (General Public License), es decir, cualquier usuario puede descargar y crear su código de manera gratuita y sin restricciones de uso. La única regla es que la distribución siempre sea libre.

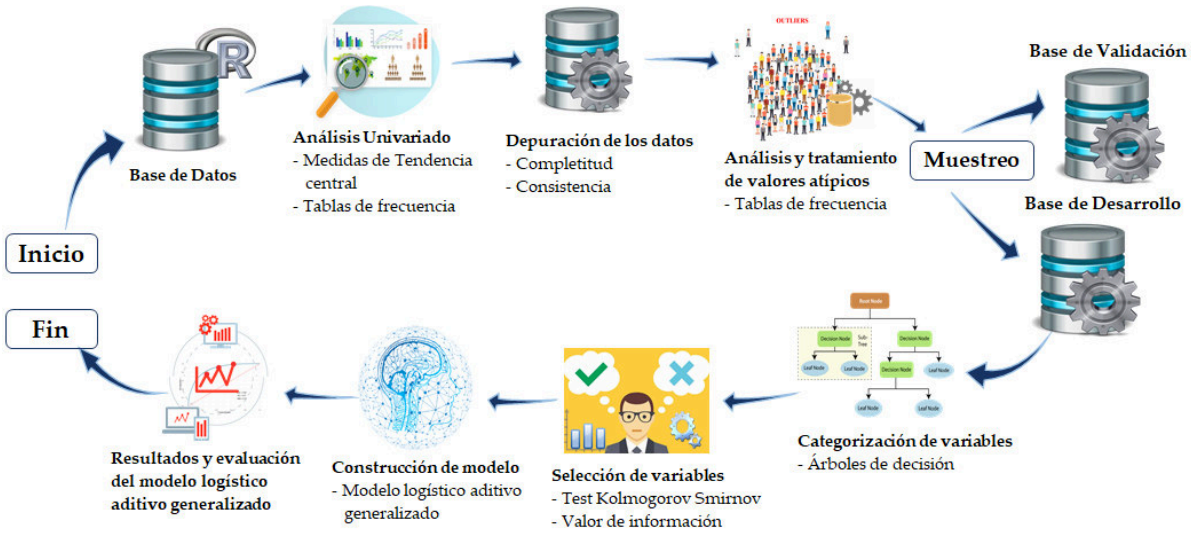
Debido a que el acceso a su código es libre, **R** no tiene limitadas su funciones, como sucede con otras herramientas estadísticas como SPSS, Statistica, etc. Lo que sin duda representa una gran ventaja.

5.2 Algoritmo implementado

El algoritmo fue implementado por completo en el software estadístico **R** con ayuda de *RStudio* (Entorno de desarrollo integrado), el cual permite explotar al máximo la capacidad que posee **R** en la ciencia de datos.

Mediante un flujograma, presentado en el gráfico de la Figura 5.1 se describen los pasos principales de la Metodología Analítica que forman parte del algoritmo.

Figura 5.1: Flujograma del algoritmo programado en **R**



Fuente: Elaboración propia.

A continuación, se describen cada uno de los pasos presentados en la Figura 5.1. Las

líneas de código del algoritmo se presentan en su totalidad y a detalle en el ANEXO 6.

5.2.1 Análisis exploratorio y tratamiento de datos

Los primeros pasos a ejecutarse corresponden a los relacionados al análisis exploratorio de los datos y tratamiento de los mismos.

5.2.1.1 Análisis Univariado

En este paso, el algoritmo ejecuta el Análisis Univariado o Estadística Descriptiva de cada una de las variables explicativas. Para las variables cuantitativas se calculan las medidas de tendencia central, posicionamiento y dispersión; y para las variables cualitativas se construyen las tablas de frecuencia.

Se presentan las funciones programadas en **R** que permiten ejecutar este paso.

- **Medidas de tendencia central, posicionamiento y dispersión**

```
#### base: Base de datos
AnalisisUniv_Cuant <- function(base){
  variable <- as.data.frame(dplyr::select_if(base, is.numeric))
  BaseCuantR <- NULL
  for (i in 1:ncol(variable)) {
    #i=1
    x <- select(variable, var=i)
    `Porc_NAs`<- round(sum(is.na(x$var))/length(x$var),4)*100
    `Porc_0s` <- round(sum(x$var=="0",na.rm=TRUE)/length(x$var),4)*100
    `Mínimo`  <- round(min(x$var,na.rm=TRUE),2)
    `Perc_25%`<- round(quantile(x$var,0.25,na.rm=TRUE),2)
    `Mediana` <- round(median(x$var,na.rm=TRUE),2)
    `Media`   <- round(mean(x$var,na.rm=TRUE),2)
    `Perc_75%`<- round(quantile(x$var,0.75,na.rm=TRUE),2)
    `Máximo`  <- round(max(x$var,na.rm=TRUE),2)
```

```
    `Desv_Est`<- round(sd(x$var,na.rm=TRUE),2)

    Aux <- tibble(`Porc_NAs`, `Porc_0s`, `Mínimo`, `Perc_25%`,
                  `Mediana`, `Media`, `Perc_75%`, `Máximo`, `Desv_Est`)
    BaseCuantR <- rbind(BaseCuantR, Aux)
  }
  BaseCuantR <- BaseCuantR %>%
  rownames_to_column("Variable") %>%
  rownames_to_column("Nº")
  BaseCuantR$Variable <- colnames(variable)
  return(BaseCuantR)
}
```

- **Tablas de frecuencia**

```
#### base: Base de datos
AnalisisUniv_Cual <- function(base) {
 base <- as.data.frame(base)
 NamesV <- names(base)
 StatM <- NULL
 for (i in 1:ncol(base)) {
   if (is.factor(base[, i]) | is.character(base[, i])) {
      StatV <- data.frame(table(base[, i], useNA = "always"))
      StatV$Var1 <- forcats::fct_explicit_na(StatV$Var1)
      if (is.null(StatM)==TRUE) {
        StatM <- data.frame(Variable = NamesV[i], StatV)
      } else {
        StatM <- rbind(StatM, data.frame(Variable = NamesV[i], StatV))
      }
    } else {
      next
    }
```

```
  }
  n <- nrow(base)
  StatM <- StatM %>%
    dplyr::rename(Variable = Variable, Categorias = Var1,
                  Frecuencia = Freq) %>%
    dplyr::group_by(Variable, Categorias, Frecuencia) %>%
    dplyr::summarise(Porcentaje = round(Frecuencia / n,4)*100) %>%
    as.data.frame()
  StatM <- data.frame(StatM)
  return(StatM)
}
```

5.2.1.2 Depuración de los datos

Luego de la exploración de los datos corresponde realizar su tratamiento, comenzando con la depuración de los mismos. Se procede como sigue:

- **Completitud:** En este paso el algoritmo imputa los valores perdidos mediante la siguiente función programada en **R**.

```
#### Imputación de variables
imputacion_variable <- function(variable, metodo = "ninguno") {
  if (any(IdNa <- is.na(variable))) {
    if (is.numeric(variable)) {
      if (metodo == "ninguno") {
        variable
      } else if (metodo == "media") {
        variable[IdNa] <- mean(variable, na.rm = TRUE)
      } else if (metodo == "mediana") {
        variable[IdNa] <- stats::median(variable, na.rm = TRUE)
      } else {
        stop("El método seleccionado no es correcto.")
      }
```

```
    } else if (is.factor(variable) | is.character(variable)) {
      if (metodo == "ninguno") {
        variable
      } else if (metodo == "moda") {
        variable[IdNa] <- levels(variable)[which.max(table(variable))]
      } else {
        stop("El método seleccionado no es correcto.")
      }
    } else {
      stop("La variable no es de tipo numeric ni de tipo factor.")
    }
  } else {
    stop("La variable no tiene valores perdidos.")
  }
  return(variable)
}

#### Imputación de variables Numéricas
psych::describe(Base$V57_Antiguedad_Laboral)
Base$V57_Antiguedad_Laboral <- imputacion_variable(
                    Base$V57_Antiguedad_Laboral, metodo = "mediana")

#### Imputación de variables categóricas
Base$V43_EsIndependiente <- imputacion_variable(
                    Base$V43_EsIndependiente, metodo = "moda")
Base$V47_TipoVivienda <- imputacion_variable(
                    Base$V47_TipoVivienda, metodo = "moda")
Base$V48_EstadoCivil <- imputacion_variable(
                    Base$V48_EstadoCivil, metodo = "moda")
Base$V50_Sector <- imputacion_variable(
                    Base$V50_Sector, metodo = "moda")
Base$V53_SectorEconomico <- imputacion_variable(
```

```
                    Base$V53_SectorEconomico, metodo = "moda")
```

- **Consistencia:** La función programada en **R** corrige las variables porcentuales que no cumplan con su regla semántica.

```
#### Análisis de Consistencia
Consistencia <- function(base, nomVar){
  for (variable in nombVar) {
    Aux <- select(base, NomVar=i)
    Aux$NomVar <- ifelse(Aux$NomVar <0, 0,  ifelse(Aux$NomVar > 1, 1,
    Aux$NomVar))
    base[,i] <- Aux
  }
  return(base)
}
```

- **Variables constantes:** En caso de existir variables constantes, la función programada elimina estas variables de la base de datos.

```
#### Análisis de variables constantes
VarConstantes <- function(base){
  variables <- dplyr::select_if(dplyr::select(base, -c(1,2,3)),
   negate(is.character))
  variables <- dplyr::select_if(variables, negate(is.factor))
  Var <- NULL
  for (i in names(variables)) {
    Aux <- dplyr::select(variables, NomVar=i)
    if(min(Aux$NomVar, na.rm = T)==max(Aux$NomVar, na.rm = T) |
        any(prop.table(table(Aux$NomVar))>=0.97)){
      Var <- c(Var, i)
    }
  }
```

```
  return(Var)
}
Var <- VarConstantes(Base)
Base <- dplyr::select(Base, -Var)
```

5.2.1.3 Análisis y tratamiento de valores atípicos

Lo que sigue es realizar el análisis y tratamiento de valores atípicos. A las variables numéricas seleccionadas se les realiza un estudio, analizando y tratando todos los valores atípicos que se encuentren. La siguiente función nos permite tratar los valores atípicos mediante el método conocido como **Winsorización**. En este estudio se hace uso de la Winsorización del $90\,\%$.

```
#### Método de Winsornización
winsorizing_method <- function(variable, removeNA = TRUE){
  percentil <- quantile(variable,
                        probs = c(.05, .95),
                        type=1, na.rm = removeNA)
  variable[variable<percentil[1]] <- percentil[1]
  variable[variable>percentil[2]] <- percentil[2]
  return(variable)
}
```

5.2.2 Selección de Muestras: Desarrollo y Validación

El siguiente paso que realiza el algoritmo es dividir la muestra total seleccionada en el mes de observación en dos submuestras: muestra de desarrollo y validación con la proporción $(70\,\%/30\,\%)$, mediante la siguiente función programada en **R**.

```
#### Muestras: Desarrollo / Validacion (70%-30%)
## Base sin Indeterminados
Base1 <- dplyr::filter(Base, Variable_Dependiente != 2)
```

```r
set.seed(123)
index <- sample(1:nrow(Base1),
                replace = FALSE,
                size = floor(0.7*nrow(Base1)))
bd_des <- Base1[index,]
bd_val <- Base1[-index,]
```

5.2.3 Categorización de variables

En este paso el algoritmo categoriza las variables cualitativas que posean 5 o más categorías y variables cuantitativas que posean un valor cuya frecuencia relativa sea mayor o igual al $20\,\%$. Esta categorización se realiza mediante árboles de decisión, con grupos de al menos el $5\,\%$ de observaciones de la muestra total en cada nodo. Se presentan las funciones programadas en **R** que permiten ejecutar este paso.

```r
#### Categorización de variables cuantitativas y cualitativas mediante
#### árboles de decisión
categorizacion <- function (base, resp, n_per = 7, n_factor = 5) {
  nombres <- colnames(base)[colnames(base) != resp]
  BDDnew_cat <- data.frame(Variable = character(0),
                           NuevaVariable = character(0),
                           Formula = character(0),
                           Detalle = character(0))
  for (i in nombres) {
    if (is.numeric(base[, i])) {
      prove_percentil <- Hmisc::cut2(base[, i], g = n_per)
      if (length(levels(prove_percentil)) != n_per) {
        regla <- extraccion_reglas_arbol(base, resp, i)
        new_nom_var <- paste(i, "cat", sep = "_")
        BDDnew_cat <- rbind(BDDnew_cat,
                            data.frame(Variable = i,
                                       NuevaVariable = new_nom_var,
```

```
                                        Formula = regla,
                                        Detalle = "pretransformacion"))
      } else {
        next()
      }
    } else {
      if (length(levels(droplevels(as.factor(base[, i])))) >= n_factor) {
        regla <- extraccion_reglas_arbol(base,
                                         respuesta = resp,
                                         variable = i)
        new_nom_var <- paste(i, "cat", sep = "_")
        BDDnew_cat <- rbind(BDDnew_cat,
                            data.frame(Variable = i,
                                       NuevaVariable = new_nom_var,
                                       Formula = regla,
                                       Detalle = "pretransformacion"))
      } else {
        next()
      }
    }
  }
  BDDnew_cat$Formula <- as.character(BDDnew_cat$Formula)
  BDDnew_cat <- dplyr::filter(BDDnew_cat, Formula != "Sin_Arbol")
  return(BDDnew_cat)
}

#### Extracción de las reglas de decisión de un árbol
extraccion_reglas_arbol <- function(base, resp, variable, n_porc = 0.05) {
  Y <- as.character(resp)
  x <- as.character(variable)
  BDD <- as.data.frame(base[, c(Y, x)])
  individuos_nodos <- round(n_porc * nrow(BDD))
```

```
formula_ctree <- formula(paste(Y, x, sep = " ~ "))
numbers_only <- function(x){
  suppressWarnings(!is.na(as.numeric(as.character(x))))
  }
try(
  if (is.numeric(BDD[, x])) {
    ct1 <- partykit::ctree(formula_ctree, data = BDD,
      control = partykit::ctree_control(minbucket = individuos_nodos))
    Nodo1 <- as.character(names(partykit:::.list.rules.party(ct1)))
    if (length(Nodo1) > 1) {
      Regla1 <- as.character(partykit:::.list.rules.party(ct1))
      sp1 <- strsplit(Regla1, split = " ")
      sp2 <- list()
      for (i in 1:length(sp1)) {
        aux <- sp1[[i]]
        sp2[[i]] <- aux[numbers_only(sp1[[i]])]
      }
      valores <- sp2
      valores_unicos <- sort(unique(as.numeric(unlist(valores))))
      new_variable <- cut(BDD[, x],
                          breaks = c(-Inf, valores_unicos, Inf),
                          ordered_result = TRUE, dig.lab = 5)
      tabla_categorias <- data.frame(Variable = x,
                            Categorias = sort(unique(new_variable)),
                            Valores = Regla1,
                            stringsAsFactors = FALSE)
      Regla2 <- NULL
      n <- length(Regla1) - 1
      for (i in 1:n) {
        Regla2[i] <- paste0("ifelse(", Regla1[i], ", ",
                            paste0("c(", "\"",
                            tabla_categorias$Categorias[i], "\"", ")"))
```

```
      }
      formula1 <- paste(Regla2, collapse = ", ")
      cierre_formula <- paste0(rep(")", n), collapse = "")
      formula2 <- paste(formula1,
                        paste0("c(", "\"",
                        tabla_categorias$Categorias[length(Regla1)],
                        "\"", ")"),  sep = ", ")
      formula_final <- paste0(formula2, cierre_formula)
    } else {
      formula_final <- NULL
    }
  } else if (is.factor(BDD[, x])) {
    BDD[, x] <- droplevels(BDD[, x])
    ct1 <- party::ctree(formula_ctree, data = BDD,
          controls = party::ctree_control(minbucket = individuos_nodos))
    aux <- data.frame(var = BDD[, x], Nodo = party::where(ct1))
    Nodos <- sort(unique(aux$Nodo))
    if (length(Nodos) > 1) {
      grupos <- list()
      for (i in 1:length(Nodos)) {
        aux_0 <- dplyr::filter(aux, Nodo == Nodos[i])
        grupos[[i]] <- sort(unique(as.character(aux_0[, "var"])))
      }
      Category <- paste("\"", paste("Grupo_", 1:length(Nodos),
                                    sep = ""), "\"", sep = "")
      tabla_categorias <- data.frame(Variable = x,
                             Categorias = Category[1],
                             Valores = paste(grupos[[1]],
                                             collapse = ", "),
                             stringsAsFactors = FALSE)
      for (i in 2:length(Category)) {
        tabla_categorias <- dplyr::bind_rows(tabla_categorias,
```

```
                                data.frame(Variable = x,
                                    Categorias = Category[i],
                                    Valores = paste(grupos[[i]],
                                                    collapse = ", "),
                                    stringsAsFactors = FALSE))
      }
      vector_grupos <- list()
      for (j in 1:length(grupos)) {
        vector_grupos[[j]] <- paste("c(", (paste0(paste("\"",
                                    grupos[[j]],"\"", sep = ""),
                                    collapse = ", ")), ")",
                                    sep = "")
      }
      n <- length(grupos) - 1
      Regla2 <- NULL
      for (i in 1:n) {
        Regla2[i] <- paste0(paste("ifelse(", x, " %in% ",
                            vector_grupos[[i]], sep = ""), ", ",
                            Category[i])
      }
      formula1 <- paste(Regla2, collapse = ", ")
      cierre <- paste0(rep(")", n), collapse = "")
      ReglaFinal <- paste(formula1, Category[length(grupos)],
                            sep = ", ")
      formula_final <- paste0(ReglaFinal, cierre)
    } else {
      formula_final <- NULL
    }
  }, silent = TRUE
)
if (is_empty(formula_final)) {
  return("Sin_Arbol")
```

```
  } else {
    return(Regla = formula_final)
  }
}
```

5.2.4 Selección de variables

En este paso el algoritmo realiza la selección de variables explicativas empleando las siguientes metodologías.

- **Medidas de separación:** La función programada en **R** calcula el valor de la prueba de Kolmogórov-Smirnov (Test KS) para cada una de las variables cuantitativas.

```
#### Estadístico Kolmogórov-Smirnov
# resp: variable dependiente.
# base: variables explicativas.
KS_test <- function(resp, base) {
  Y <- base[, resp]
  variable <- as.data.frame(select_if(base[, setdiff(names(base),
                                                     resp)],is.numeric))
  Aux <- NULL
  for (i in 1:ncol(variable)) {
    m <- data.frame(Y, variable[,i])
    m1 <- filter(m, Y == "1")
    m2 <- filter(m, Y == "0")
    ks <- suppressWarnings(
      stats::ks.test(
        m1[, 2],
        m2[, 2],
        alternative = "two.sided",
        exact = FALSE
      )
    )
```

```
    Aux1 <- data.frame(Valor_KS =round(as.numeric(ks$statistic), 4))
    Aux <- rbind(Aux, Aux1)
  }
  Tabla_KS <- tibble::rownames_to_column(Aux, "Variable")
  Tabla_KS$Variable  <- names(variable)
  Tabla_KS <- dplyr::arrange(Tabla_KS, desc(Valor_KS))
  return(Tabla_KS)
}
```

- **Medidas de asociación:** La función programada en **R** calcula el valor de información (VI) para cada una de las variables cualitativas.

```
#### Calcula el estadístico de Information-Value
# resp: variable dependiente.
# base: variables explicativas.
Information_Value <- function(resp, base){
  Y <- base[, resp]
  variable <- as.data.frame(select_if(base[, setdiff(names(base),
                                                     resp)], is.factor))
  IV <- NULL
  for (i in 1:ncol(variable)) {
    Frec <- table(Y,variable[,i])
    Frec1 <- Frec[1,]
    Frec2 <- Frec[2,]
    Aux1 <- ifelse(Frec1/sum(Frec1)==0, 0.0001,
                   ifelse(Frec1/sum(Frec1)==1, 0.999,
                          Frec1/sum(Frec1)))
    Aux2  <- ifelse(Frec2/sum(Frec2)==0, 0.0001,
                    ifelse(Frec2/sum(Frec2)==1, 0.999,
                           Frec2/sum(Frec2)))
    Woe <- log(Aux2/Aux1)
    Woe <- ifelse(Woe==-Inf, 0, Woe)
```

```
    IV1 <- data.frame(Valor_IV = sum((Frec2/sum(Frec2) -
                                      Frec1/sum(Frec1))*Woe))
    IV <- rbind(IV, IV1)
  }
  Tabla_IV <- tibble::rownames_to_column(IV, "Variable")
  Tabla_IV$Variable  <- names(variable)
  Tabla_IV <- dplyr::arrange(Tabla_IV, desc(Valor_IV))
  return(Tabla_IV)
}
```

5.2.5 Construcción del modelo logístico aditivo generalizado

El modelo logístico aditivo generalizado óptimo se lo construye suavizando las variables explicativas cuantitativas mediante splines cúbicos de regresión y el número de nodos adecuado. Se presenta la función en **R** que permite construir este modelo.

```
#### Modelo logístico aditivo generalizado óptimo
log_mod <- gam(Variable_Dependiente ~
                 ##### Numéricas
                 V16_Amortizacion +
                 #s(V16_Amortizacion, bs = 'cr') +
                 #V28_MaxAtraso_DC +
                 #V31_AtrasoMax_U6_AC +
                 #s(V31_AtrasoMax_U6_AC, bs = 'cr') +
                 s(V31_AtrasoMax_U6_AC, bs = 'cr', k = 12) +
                 #V32_AtrasoMax_U12_AC +
                 #V33_AtrasoMax_U18_AC +
                 #V36_MontoLiquido_Total +
                 V38_AntiguedadCliente +
                 #s(V38_AntiguedadCliente, bs = 'cr') +
                 #V44_AntiguedadResidencia +
                 V46_Edad +
```

```
                #s(V46_Edad, bs = 'cr') +
                V60_NumeroInst_Adeuda +
                #s(V60_NumeroInst_Adeuda, bs = 'cr', k = 5) +
                #V61_NumeroInst_Adeuda_3 +
                #V62_NumeroInst_Adeuda_6 +
                #V68_TiempoHistorialCrediticioSF_12 +
                #V70_Num_Calificacion_0 +
                #V81_Amortizacion_comp +
                #s(V81_Amortizacion_comp, bs = 'cr') +
                s(V81_Amortizacion_comp, bs = 'cr', k =6) +
                V90_VariacionDeudaMicrocredito_U6M +
                #s(V90_VariacionDeudaMicrocredito_U6M, bs = 'cr') +
                V91_VariacionDeudaTarjeta_U6M +
                #s(V91_VariacionDeudaTarjeta_U6M, bs = 'cr') +
                ##### Categóricas
                V17_CarteraRiesgo_Q +
                V18_SaldoMMora_Mont +
                #V25_Porc_CuotasPag +
                #V26_Porc_CuotasVenc +
                V29_PromAtraso_DC +
                #V71_Num_Calificacion_1 +
                #V72_Num_Calificacion_234 +
                #V73_PeorCalificacionCorte_Comp +
                #V74_PeorCalificacionU6M_Comp +
                #V75_PeorCalificacionU12M_Comp +
                #V76_PeorCalificacionU18M_Comp +
                #V77_CarteraRiesgoPond_U6M_Comp +
                #V78_CarteraRiesgo_comp +
                V79_MoraPonderada_comp +
                V82_Saldo_MB_Comp,
              data = bd_train,
              family = binomial,
```

```
                method = "REML",
                select = TRUE
                )
```

5.2.6 Resultados y evaluación del modelo logístico aditivo generalizado

Finalmente, el algoritmo calcula los estadísticos: AUC, KS y coeficiente de GINI (Medidas de discriminación), la matriz de confusión y la tabla de desempeño. Estas técnicas estadísticas nos ayudan a evaluar la capacidad predictiva y de discriminación del modelo, mediante las siguientes funciones programadas en **R**.

```
#### Medidas de discriminación (AUC, KS, GINI)
MLmetrics::AUC(prediccion, respuesta)
MLmetrics::KS_Stat(prediccion, respuesta)/100
MLmetrics::Gini(prediccion, respuesta)

### Matriz de confusión
caret::confusionMatrix(prediccion, respuesta, positive = '1')

### Tabla de desempeño
Tabla_Performance <- function(prediccion, respuesta, n = 10){
  Performance <- data.frame(prediccion, respuesta) %>%
    dplyr::mutate(respuesta = as.numeric(as.character(respuesta)),
                  Deciles = factor(Hmisc::cut2(prediccion, g = n,
                                               digits = 3))) %>%
    dplyr::group_by(Deciles) %>%
    dplyr::summarise(N = n(), Porc_N = round(n()/length(obj),4)*100) %>%
    dplyr::mutate(Acum_N=cumsum(Porc_N))
  Aux <- as.data.frame.matrix(table(Hmisc::cut2(prediccion,
                                                g = n,
                                                digits = 3),
```

```
                                    respuesta)) %>%
    rownames_to_column("Deciles")
  Performance <- data.frame(Performance, Buenos=Aux[,3],
                            row.names=NULL) %>%
    dplyr::mutate(Porc_Buenos = round(Buenos/sum(Buenos),3)*100) %>%
    dplyr::mutate(Acum_Buenos = cumsum(Porc_Buenos))
  Performance <- data.frame(Performance, Malos=Aux[,2],
                            row.names=NULL) %>%
    dplyr::mutate(Porc_Malos = round(Malos/sum(Malos),3)*100) %>%
    dplyr::mutate(Acum_Malos = cumsum(Porc_Malos)) %>%
    dplyr::mutate(Razon_Buenos = round(Buenos/N,5)*100) %>%
    dplyr::mutate(Razon_Malos = round(Malos/N,3)*100)
  return(Performance)
}
```

Capítulo 6

Conclusiones y Recomendaciones

En la actualidad tanto instituciones financieras como comerciales cuentan con herramientas estadísticas que presentan dificultades al momento de modelar características no lineales, por lo que los resultados obtenidos podrían mejorarse con el uso de una herramienta estadística moderna y confiable que además de capturar esta no linealidad presente tanto en información del solicitante de crédito, interna de la institución y de la central de riesgos; cree un modelo robusto que pueda ser interpretado fácilmente, apoyando así en la toma de decisiones que permitan mitigar el riesgo de incumplimiento asociado a la concesión de crédito, logrando con ello mantener una liquidez y solvencia adecuadas. Esta alternativa puede resultar prometedora para la calificación crediticia.

El presente trabajo tiene como finalidad comparar empiricamente si emplear modelos aditivos generalizados, para estimar la probabilidad de incumplimiento de un cliente al momento de la concesión del crédito, logran un mejor desempeño que emplear una metodología tradicional como la regresión logística. Esta herramienta nos permitirá presentar un modelo estadístico robusto y con ello establecer los determinantes del incumplimiento del crédito de tal forma que la entidad financiera pueda llevar a cabo planes de acción para manejar adecuadamente su cartera.

La Base de datos que se empleó en este estudio, corresponde a la cartera de créditos concedidos de a una institución financiera de un país emergente. La misma que dispone de información demográfica, interna (describe el historial crediticio del cliente

dentro de la institución) y externa (describe el comportamiento del cliente en instituciones externas). La naturaleza de la información proporcionada nos permite construir modelos scoring de comportamiento.

A partir de la información proporcionada, se construyeron tres modelos de scoring crediticio, un modelo bajo la metodología de modelos aditivos generalizados $(GAM-2)$, una regresión logística con las variables brutas (o transformadas si estas tienen una relación no lineal con la variable dependiente) que ingresaron al modelo aditivo generalizado $(RL-M1)$ y una regresión logística tradicional $(RL-M2)$. Al comparar los tres modelos se determinó que el modelo $(GAM-2)$ logra una superioridad (aunque mínima) sobre el resto de modelos; y el modelo que más se aproxima es el modelo $(RL-M1)$ que se construye empleando las variables que forman parte del modelo $(GAM-2)$. Es importante señalar que el modelo $(GAM-2)$ emplea funciones suaves para modelar las relaciones no lineales que puedan tener ciertas variables cuantitativas con la variable dependiente. Las variables en las que se empleó funciones suaves y con las que se obtuvo un modelo GAM óptimo fueron: el máximo atraso de la persona 6 meses atrás hasta la fecha de corte ($V31\_AtrasoMax\_U6\_AC$) y la amortización del producto de crédito, de la competencia ($V81\_Amortizacion\_comp$). El resto de variables cuantitativas no fueron suavizadas debido a que no resultaban ser significativas, por lo que, a estas no se les realiza ninguna acción.

Cabe recalcar que entre los principales aportes que realiza este trabajo se encuentra la descripción detallada de cada uno de los pasos a seguir para la construcción de los modelos scoring de comportamiento (metodología analítica), la construcción de un modelo estadístico robusto que con ayuda de modelos aditivos generalizados obtenga un poder predictivo alto sin sacrificar la interpretabilidad. Por último, la implementación del algoritmo creado a partir de la metodología analítica en el software estadístico **R**. El cual ejecuta de manera automática cada uno de los pasos de la metodología desarrollada.

A continuación, se exponen las principales **conclusiones** elaboradas en virtud del estudio realizado:

1. Con ayuda de la metodología analítica se construyó el mejor modelo logístico aditivo generalizado $(GAM-2)$. El cual emplea splines cúbicos de regresión para suavizar las variables cuantitativas no lineales. Además, se eligieron 12 nodos para la variable $V31\_AtrasoMax\_U6\_AC$ y 6 nodos para la variable $V81\_Amortizacion\_comp$, estos nodos dividen a la variable en 13 y 7 secciones respectivamente y en cada sección se ajusta una spline cúbica de regresión (polinomio de grado 3).

2. El modelo logístico aditivo generalizado construido nos permite evidenciar las principales causas de que un cliente sea $(Bueno/Malo)$ mediante las siguientes variables:

    - Amotización del producto de crédito de la competencia: Si un cliente presenta mora en otras instituciones financieras, existe una alta probabilidad de que sea moroso en nuestra institución. Estos clientes tienen menos probabilidad de ser un Buen pagador.
    - Monto del máximo atraso: Clientes que tienen el máximo valor en atraso alto en los últimos 6 meses tienen menor probabilidad de ser Buenos pagadores.
    - Atraso promedio en meses: Tienen menor probabilidad de ser Buen pagador aquellos clientes que presentan en promedio más meses en atraso.
    - Antigüedad del cliente en la Institución Financiera: Un cliente que tienen mayor antigüedad en la institución, tienen mayor probabilidad de ser un Buen pagador.
    - Número de instituciones en las que adeuda: Mientras mayor sea el número de institituciones en las que el cliente tiene deuda menor es la probabilidad de que sea un Buen pagador.

3. Los modelos aditivos generalizados nos permiten construir un modelo con un mejor estadístico de Kolmogórov-Smirnov (KS), índice AUC y coeficiente de Gini (GINI) que los modelos que emplean una regresión logística bajo el primer método $(RL-M1)$ y bajo el segundo método $(RL-M2)$, (Método tradicional) (Tabla 4.13). Es decir, existe una mayor discriminación entre clientes Buenos y

Malos con lo cual se obtiene una clasificación más precisa. La diferencia radica en que los modelos aditivos generalizados modelan las relaciones no lineales que poseen las variables cuantitativas con la variable dependiente.

4. Analizando el ordenamiento (posible segmentación de los clientes) en las tablas de desempeño del modelo logístico aditivo generalizado $(GAM-2)$ (Tabla 3.15), del modelo $(RL-M1)$ (Tabla 4.6) y del modelo $(RL-M2)$ (Tabla 4.11), podemos concluir que los tres modelos son adecuados ya que la tasa de clientes Buenos en cada intervalo de probabilidad aumenta cuando la probabilidad aumenta y lo hace con un crecimiento estricto, capturando porcentajes mayores en los deciles más altos. Sin embargo, la tasa de clientes Buenos del modelo $(GAM-2)$ (Figura 4.7) presenta un ordenamiento más parsimonioso, por lo que segmenta de mejor manera a los clientes Buenos. Por otro lado, el ordenamiento del modelo $(RL-M1)$ es el que más se aproxima al ordenamiento del modelo $(GAM-2)$.

5. El software estadístico **R**, al ser un lenguaje y un ambiente de programación libre, intuitivo y de fácil implementación, nos permite programar el algoritmo desarrollado a partir de la metodología analítica (Capítulo 3). El flujograma de la Figura 5.1 del capítulo 5 resume cada uno de los pasos del algoritmo implementado. En el ANEXO 6 se presenta en su totalidad las líneas de código del algoritmo.

6. En la práctica ocurre que para capturar la no linealidad presente en modelos de regresión mediante alguna función f, el experto transforma o categoriza alguno o todos los predictores y así poder modelar esta no linealidad. Sin embargo, al transformar matemáticamente una variable se puede perder la relación real existente entre la variable respuesta y las variables explicativas, ya que pueden existir relaciones que tengan una forma desconocida, el modelo logístico aditivo generalizado permite modelar de manera flexible las relaciones no lineales sin realizar ninguna suposición sobre la forma funcional de f, esto ofrece una mejor predicción sin perder su capacidad interpretativa. Por lo que, usar un modelo GAM puede resultar más sencillo.

7. Lo ideal es obtener reglas explícitas que determinen la probabilidad de buen pagador de un cliente, los modelos aditivos generalizados presentan una difi-

cultad en este punto, ya que los coeficientes de términos no paramétricos, son coeficientes de una forma funcional de la variable, mas no, coeficientes de las variables. En otras palabras, un término no paramétrico puede contar con n coeficientes, dependiendo del número funciones básicas empleadas para modelar la no linealidad de una variable (n número de nodos determina n-1 funciones básicas). Por tal motivo el modelo $(RL - M1)$, regresión logística construida a partir del mismo conjunto de variables que ingresan al modelo $(GAM - 2)$ y el que más se aproxima, se puede considerar como una buena opción cuando el negocio desee obtener reglas explícitas que determinen la probabilidad de buen pagador.

8. El conjunto final de variables que ingresan al modelo logístico aditivo generalizado $(GAM - 2)$ es diferente al conjunto de variables que ingresan al modelo de regresión logística **tradicional** $(RL - M2)$, lo que nos permite concluir que los modelos aditivos generalizados son una herramienta útil para realizar selección de variables cuantitativas, ya que se obtienen mejores resultados al momento de perfilar a los clientes.

9. La calidad de discriminación de un modelo, está dado por la calidad de información que se disponga. Las medidas de separación y asociación son técnicas útiles, para construir criterios de selección de variables lo que permite mejorar el rendimiento de discriminación del modelo y disminuir considerablemente el tiempo de desarrollo.

Es útil contar con recomendaciones que sirvan de apoyo en trabajos futuros. Por lo cual, se realizan las siguientes recomendaciones a partir del desarrollo de los modelos estadísticos.

1. Es usual emplear la regresión logística tradicional en la construcción de modelos $Credit\ Score$ ya que se entiende relativamente bien y se puede derivar una fórmula explícita en la que se puedan basar las decisiones de crédito. Sin embargo, en la actualidad existen metodologías como los modelos aditivos generalizados que logran mejores predicciones y aún se pueden comprender y explicar por qué hacen las predicciones que hacen. Metodologías no tradicionales que se recomienda sean utilizadas en la construcción de modelos estadísticos robustos.

2. A pesar del "Boom"que existe en la actualidad por el Big Data y Analítica de datos, ciertas instituciones financieras no cuentan con bases de datos con información histórica necesaria que describa las características de los clientes, que sea confiable y actualizada. Por lo que, enriquecer las bases de datos que poseen es una buena práctica para obtener modelos estadísticos con mayor poder predictivo y de discriminación.

3. En estudios futuros se podría profundizar en el uso de modelos aditivos generalizados orientados a la categorización de variables cuantitativas para que esta categorización pueda ser empleada en la construcción de modelos estadísticos bajo diferentes técnicas.

4. Dado que en este estudio se centró en la construcción de modelos, que además de tener un poder predictivo alto puedan ser interpretados, las funciones suaves empleadas fueron los splines cúbicos de regresión. Sin embargo, en trabajos futuros, si lo que interesa es obtener un mayor poder de predicción, se pueden emplear distintas funciones suaves disponibles que permiten modelar incluso interacciones entre variables predictoras no lineales (Por ejemplo, splines de regresión Thin Plate) y con ello evaluar los diferentes resultados.

5. Realizar monitoreo y calibración del modelo construido periódicamente. Se recomienda realizarlo cada 3 meses con la finalidad de analizar su poder de predicción y estabilidad a lo largo del tiempo.

Bibliografía

[Anderson, 2007] Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press Inc, USA, New York.

[Superintendencia de Bancos y Seguros, Libro 1] SUPERINTENDENCIA DE BANCOS DEL ECUADOR. *Libro Normas generales para título IX. De los activos y de las instituciones del sistema y de los límites crédito*. Quito.

[Aguilar y Camargo, 2004] Aguilar Andía, A.G. y Camargo, C.G. (2004). *Instituto de Estudios peruanos. Serie Economía. Recuperado el 11 de 10 de 2019*, de: http://lanic.utexas.edu/project/laoap/iep/ddt133.pdf

[González y López, 2008] González, A., & López, J. (2008). *Gestión Bancaria. Factores claves en un entorno competitivo*. Madrid, España McGraw-Hill / Interamericana de España. S.A.U.

[Franke, Hardle and Stahl, 2008] J. Franke, W. Härdle and G. Stahl (2000). *Measuring Risk in Complex Stochastic Systems*. Springer Verlag.

[Shao, 2004] Shao, N. (2004). *Semi-Parametric Estimation for Credit Scoring*, Research Projects, Department of Quantitative Analysis, University of Cincinnati.

[Elizondo, 2003] Elizondo, A. (2003). *Medición Integral del Riesgo de Crédito*. México. Editorial Limusa. 21

[Liu and Cela, 2007] Liu, W. and Cela, J. (2007). *Improving Credit Scoring by Generalized Additive Model*. SAS Institute Inc., 078,1-14.

[Hastie and Tibshirani, 1990] Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*. London: Chapman & Hall/CRC.

[Girault, 2007] Girault, M. (2007). *Modelos de credit scoring:¿ Qué, cómo, cuándo, y para qué?*. Buenos Aires, Argentina: Gerencia de Investigación y Planificación Normativa, Subgerencia General de Normas, BCRA.

[Kleinbaum, 1994] Kleinbaum, D. G. (1994). *Statistics in the health sciences: logistic regression*. hird Edition, New York: Springer-Verlag

[Ross, 2019] Ross, N. (2019). *Generalized Additive Models (GAMs) in R*. URL: https://noamross.github.io/gams-in-r-course/

[Liu and Chuck, 2009] Liu, W. and Chuck, C. (2009). *Generalizations of Generalized Additive Model (GAM): A Case of Credit Risk Modeling*. SAS Institute Inc., 113,1-9.

[Lohmann and Ohliger, 2018] Lohmann, C. and Ohliger, T. (2018). *Nonlinear relationships in a logistic model of default for a high-default installment portfolio*. Journal of Credit Risk, 14(1),45-68.

[Kraus, 2014] Kraus, A. (2014). *Recent Methods from Statistics and Machine Learning for Credit Scoring.* Múnich, Alemania.

[Arnold and Emerson, 2011] Arnold, T. and Emerson, J. (2011). *Nonparametric goodness-of-fit tests for discrete null distributions. R Journal*, 3:34-35.

[Finlay, 2010] Finlay, S. (2010). *Credit Scoring, Response Modelling and Insurance Rating: A Practical Guide to Forecasting Consumer Behaviour*. Palgrave Macmillan, New York.

[Siddiqi, 2006] Siddiqi, Naeem. 2006. *Credit risk scorecards: developing and implementing intelligent credit scoring.*. Hoboken, N.J: Wiley.

[James et al., 2014] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in R*. Springer Publishing Company.

[Allison, P. D., 2012] Allison, P. D. (2014). *Logistic regression using SAS: theory and application.* 2nd ed. Cary, N.C: SAS Pub.

[Gujarati and Porter, 2010] Gujarati, Damodar N. & Porter, Dawn C. (2010). *Econometría*. 5ta ed. México: McGraw-Hill.

[Wood, Simon N., 2017] Wood, Simon N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. London: Chapman & Hall/CRC Press.

[Gu, 2002] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

ANEXO 1: Descripción de variables explicativas

Nº	Variable	Descripción
1	V11_TieneGarante	Indica si el cliente tiene Garante.
2	V16_Amortizacion	Saldo / MB.
3	V17_CarteraRiesgo_Q	SaldoVencido / Saldo.
4	V18_SaldoMMora_Mont	Saldo Mora/ Monto
5	V19_SaldoMoraM_Mont	Saldo Mora/ Monto
6	V25_Porc_CuotasPag	Porcentaje de cuotas pagadas.
7	V26_Porc_CuotasVenc	Porcentaje de cuotas vencidas.
8	V28_MaxAtraso_DC	Máximo atraso de la operación desde el desembolso hasta el corte.
9	V29_PromAtraso_DC	Atraso promedio de la operación desde el desembolso hasta el corte.
10	V31_AtrasoMax_U6_AC	Máximo atraso de la persona T6 meses atrás hasta la fecha de corte(T0).
11	V32_AtrasoMax_U12_AC	Máximo atraso de la persona T12 meses atrás hasta la fecha de corte(T0).
12	V33_AtrasoMax_U18_AC	Máximo atraso de la persona T18 meses atrás hasta la fecha de corte(T0).
13	V34_Nop_Total	Número de operaciones totales.
14	V34_Nop_Total_v2_M8	Número de operaciones totales en los últimos 8 meses antes del corte.
15	V35_Nop_Retanqueo_Total	Número de operaciones totales con retanqueo (renegociacion).
16	V36_MontoLiquido_Total	Monto líquido total.
17	V38_AntiguedadCliente	Diferencia en meses entre la fecha del primer desembolso y la fecha de corte.
18	V39_CargasFamiliares	Indica el número de cargas familiares.
19	V43_EsIndependiente	Indica si el cliente es Dependiente o Independiente.
20	V44_AntiguedadResidencia	Indica la antigüedad de la residencia (en años).
21	V45_Genero	Indica el género del cliente (M/F).
22	V46_Edad	Indica la edad del cliente al corte (en años).
23	V47_TipoVivienda	Indica el tipo de vivienda del cliente.
24	V48_EstadoCivil	Indica el estado civil del cliente.
25	V49_NivelEstudios	Nivel de estudios .
26	V50_Sector	Sector de vivienda.
27	V51_TieneTelefono	Tiene teléfono.
28	V53_SectorEconomico	Sector económico al que pertenece la actividad que desarrolla.
29	V56_Departamento_Dom	Ciudad de domicilio.
30	V57_Antiguedad_Laboral	Antigüedad laboral (en años).
31	V58_Antiguedad_SUNAT	

Nº	Variable	Descripción
32	V59_Auto	Indica si el cliente tiene auto.
33	V60_NumeroInst_Adeuda	Indica el número de instituciones en las que el cliente tiene deuda en la central de Riesgos en la fecha de corte.
34	V61_NumeroInst_Adeuda_3	Indica el número de instituciones en las que el cliente tiene deuda en la central de Riesgos en la fecha de corte. Con rezago 3.
35	V62_NumeroInst_Adeuda_6	Indica el número de instituciones en las que el cliente tiene deuda en la central de Riesgos en la fecha de corte. Con rezago 6.
36	V63_TieneCreditoHipotecario	Tiene crédito hipotecario.
37	V65_TiempoHistorialCrediticioSF_36	Número de meses que el sujeto tiene historial crediticio contar meses donde se reporten saldos de operaciones. Mirando 36 meses antes de la fecha de corte.
38	V66_TiempoHistorialCrediticioSF_24	Número de meses que el sujeto tiene historial crediticio contar meses donde se reporten saldos de operaciones. Mirando 24 meses antes de la fecha de corte.
39	V67_TiempoHistorialCrediticioSF_18	Número de meses que el sujeto tiene historial crediticio contar meses donde se reporten saldos de operaciones. Mirando 18 meses antes de la fecha de corte.
40	V68_TiempoHistorialCrediticioSF_12	Número de meses que el sujeto tiene historial crediticio contar meses donde se reporten saldos de operaciones. Mirando 12 meses antes de la fecha de corte.
41	V69_TiempoHistorialCrediticioSF_6	Número de meses que el sujeto tiene historial crediticio contar meses donde se reporten saldos de operaciones. Mirando 6 meses antes de la fecha de corte.
42	V70_Num_Calificacion_0	No. de veces que el cliente tiene deuda con peor calificación mensual "0.en la central de Riesgos en los U12M.
43	V71_Num_Calificacion_1	No. de veces que el cliente tiene deuda con peor calificación mensual "1.en la central de Riesgos en los U12M.
44	V72_Num_Calificacion_234	No. de veces que el cliente tiene deuda con peor calificación mensual "234.en la central de Riesgos en los U12M.
45	V73_PeorCalificacionCorte_Comp	Se toma el máximo para aquellos valores donde el saldo es mayor a cero, de la competencia.
46	V74_PeorCalificacionU6M_Comp	Se toma el máximo para aquellos valores donde el saldo es mayor a cero. Mirando 6 meses de histórico
47	V75_PeorCalificacionU12M_Comp	Se toma el máximo para aquellos valores donde el saldo es mayor a cero. Mirando 12 meses de histórico
48	V76_PeorCalificacionU18M_Comp	Se toma el máximo para aquellos valores donde el saldo es mayor a cero. Mirando 18 meses de histórico
49	V77_CarteraRiesgoPond_U6M_Comp	Cartera en riesgo ponderada 6M sobre el total de la fecha corte.
50	V78_CarteraRiesgo_comp	Cartera en riesgo en la fecha de corte solo en las otras instituciones.
51	V79_MoraPonderada_comp	Edad de Mora ponderada por el valor de cartera en cada corte.
52	V80_LineaNoUtilizada_Comp	Promedio lineal del valor de linea no utilizado en los ultimos 6 meses por producto.
53	V81_Amortizacion_comp	Amortización del producto de drédito, de la competencia.
54	V82_Saldo_MB_Comp	Saldo en atraso (vencido + catigado) en la competencia al corte / MB_Qapaq al corte.

Nº	Variable	Descripción
55	V85_VariacionDeudaConsumo_U3M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U3M.
56	V86_VariacionDeudaMicrocredito_U3M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U3M.
57	V87_VariacionDeudaTarjeta_U3M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U3M.
58	V88_VariacionDeudaTotal_U3M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U3M.
59	V89_VariacionDeudaConsumo_U6M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U6M.
60	V90_VariacionDeudaMicrocredito_U6M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U6M.
61	V91_VariacionDeudaTarjeta_U6M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U6M.
62	V92_VariacionDeudaTotal_U6M	El porcentaje de variación de deuda entre la central disponible al corte de análisis y el corte de los U6M.
63	V100_TieneTarjetaCredito	Tiene tarjeta de crédito.

ANEXO 2: Análisis Univariado

Medidas de Tendencia Central, Posicionamiento y Dispersión

Nº	Variable	Porc_NAs	Porc_0s	Mínimo	Perc_25 %	Mediana	Media	Perc_75 %	Máximo	Desv_Est
1	V16_Amortizacion	0,00 %	0,00 %	0,00	0,41	0,58	0,54	0,71	0,95	0,20
2	V17_CarteraRiesgo_Q	0,00 %	87,36 %	0,00	0,00	0,00	0,02	0,00	1,00	0,08
3	V18_SaldoMMora_Mont	0,00 %	75,54 %	0,00	0,00	0,00	0,01	0,00	0,55	0,03
4	V19_SaldoMoraM_Mont	0,00 %	75,54 %	0,00	0,00	0,00	0,01	0,00	0,55	0,03
5	V25_Porc_CuotasPag	0,00 %	0,00 %	0,80	1,00	1,00	0,99	1,00	2,67	0,05
6	V26_Porc_CuotasVenc	0,00 %	87,36 %	0,00	0,00	0,00	0,01	0,00	0,20	0,04
7	V28_MaxAtraso_DC	0,00 %	13,46 %	0,00	1,00	4,00	8,74	11,00	201,00	12,69
8	V29_PromAtraso_DC	0,00 %	48,52 %	0,00	0,00	1,00	2,10	3,00	95,00	3,97
9	V31_AtrasoMax_U6_AC	0,00 %	15,48 %	0,00	1,00	4,00	8,26	11,00	316,00	12,38
10	V32_AtrasoMax_U12_AC	0,00 %	12,74 %	0,00	1,00	4,00	8,85	11,00	316,00	12,84
11	V33_AtrasoMax_U18_AC	0,00 %	12,05 %	0,00	2,00	4,00	8,98	12,00	316,00	12,86
12	V34_Nop_Total	0,00 %	0,00 %	1,00	1,00	1,00	1,30	1,00	9,00	0,68
13	V34_Nop_Total_v2_M8	0,00 %	0,00 %	1,00	1,00	1,00	1,30	1,00	8,00	0,68
14	V36_MontoLiquido_Total	0,00 %	0,00 %	500,00	2.417,36	3.000,00	3.925,38	4.500,00	111.000,00	2.743,16
15	V38_AntiguedadCliente	0,00 %	0,00 %	3,00	7,00	10,00	13,28	15,00	86,00	9,64
16	V39_CargasFamiliares	0,00 %	40,73 %	0,00	0,00	1,00	0,92	1,00	10,00	0,98
17	V44_AntiguedadResidencia	0,00 %	1,41 %	0,00	8,00	16,00	18,96	27,00	117,00	13,39
18	V46_Edad	0,00 %	0,00 %	21,00	34,00	42,00	43,24	52,00	73,00	11,67
19	V57_Antiguedad_Laboral	0,02 %	59,72 %	0,00	0,00	0,00	1,52	1,00	102,00	4,48
20	V58_Antiguedad_SUNAT	78,66 %	0,01 %	0,00	6,00	9,00	10,46	14,00	24,00	4,91
21	V60_NumeroInst_Adeuda	0,00 %	0,00 %	1,00	2,00	3,00	2,91	4,00	9,00	1,15
22	V61_NumeroInst_Adeuda_3	0,00 %	0,00 %	1,00	2,00	3,00	2,90	4,00	9,00	1,15
23	V62_NumeroInst_Adeuda_6	0,00 %	0,03 %	0,00	2,00	3,00	2,85	4,00	8,00	1,13
24	V65_TiempoHistorialCrediticioSF_36	0,00 %	0,00 %	5,00	27,00	35,00	31,32	36,00	36,00	6,38
25	V66_TiempoHistorialCrediticioSF_24	0,00 %	0,00 %	5,00	23,00	24,00	22,70	24,00	24,00	2,67
26	V67_TiempoHistorialCrediticioSF_18	0,00 %	0,00 %	5,00	18,00	18,00	17,56	18,00	18,00	1,40
27	V68_TiempoHistorialCrediticioSF_12	0,00 %	0,00 %	5,00	12,00	12,00	11,88	12,00	12,00	0,62
28	V70_Num_Calificacion_0	0,00 %	3,03 %	0,00	9,00	12,00	10,03	12,00	12,00	3,12
29	V71_Num_Calificacion_1	0,00 %	70,57 %	0,00	0,00	0,00	0,58	1,00	12,00	1,18
30	V72_Num_Calificacion_234	0,00 %	86,73 %	0,00	0,00	0,00	0,48	0,00	12,00	1,62
31	V73_PeorCalificacionCorte_Comp	0,00 %	74,97 %	-1,00	0,00	0,00	0,26	0,00	4,00	1,00
32	V74_PeorCalificacionU6M_Comp	0,00 %	67,44 %	-1,00	0,00	0,00	0,46	1,00	4,00	1,06

Nº	Variable	Porc_NAs	Porc_0s	Mínimo	Perc_25 %	Mediana	Media	Perc_75 %	Máximo	Desv_Est
33	V75_PeorCalificacionU12M_Comp	0,00 %	66,09 %	-1,00	0,00	0,00	0,54	1,00	4,00	1,05
34	V76_PeorCalificacionU18M_Comp	0,00 %	65,68 %	-1,00	0,00	0,00	0,57	1,00	4,00	1,05
35	V77_CarteraRiesgoPond_U6M_Comp	0,00 %	83,08 %	-1,00	0,00	0,00	-0,02	0,00	1,00	0,26
36	V78_CarteraRiesgo_comp	0,00 %	83,10 %	-1,00	0,00	0,00	-0,05	0,00	1,00	0,34
37	V79_MoraPonderada_comp	0,00 %	58,96 %	-1,00	0,00	0,00	11,17	2,86	6.323,67	85,86
38	V80_LineaNoUtilizada_Comp	0,00 %	33,40 %	0,00	0,00	480,04	1.933,93	2.247,40	97.370,29	4.038,10
39	V81_Amortizacion_comp	0,00 %	8,57 %	0,00	0,52	1,33	2,51	3,03	273,74	4,21
40	V82_Saldo_MB_Comp	0,00 %	91,61 %	0,00	0,00	0,00	0,08	0,00	254,31	0,94
41	V85_VariacionDeudaConsumo_U3M	0,00 %	0,24 %	-0,99	-0,21	-0,13	-0,07	-0,08	60,29	0,74
42	V86_VariacionDeudaMicrocredito_U3M	0,00 %	62,78 %	-1,00	-0,08	0,00	0,09	0,00	79,62	1,32
43	V87_VariacionDeudaTarjeta_U3M	0,00 %	53,43 %	-1,00	-0,03	0,00	0,11	0,00	103,33	1,25
44	V88_VariacionDeudaTotal_U3M	0,00 %	0,11 %	-0,99	-0,19	-0,11	0,01	-0,01	38,78	0,65
45	V89_VariacionDeudaConsumo_U6M	0,00 %	0,01 %	-0,99	-0,38	-0,26	-0,13	-0,15	74,91	0,84
46	V90_VariacionDeudaMicrocredito_U6M	0,00 %	59,00 %	-1,00	-0,01	0,00	0,16	0,00	84,61	1,64
47	V91_VariacionDeudaTarjeta_U6M	0,00 %	50,76 %	-1,00	0,00	0,00	0,27	0,00	106,50	2,00
48	V92_VariacionDeudaTotal_U6M	0,00 %	0,02 %	-0,99	-0,32	-0,16	0,02	0,10	74,91	0,83

Tablas de Frecencias

Nº	V11_TieneGarante	Frecuencia	Porcentaje
1	0	127.413	100,00 %
2	(Missing)	0	0,00 %
3	Total	127.413	100,00 %

Nº	V35_Nop_Retanqueo_Total	Frecuencia	Porcentaje
1	0	111.135	87,22 %
2	1	12.943	10,16 %
3	2	2.576	2,02 %
4	3	657	0,52 %
5	4	102	0,08 %
6	(Missing)	0	0,00 %
7	Total	127.413	100,00 %

Nº	V45_Genero	Frecuencia	Porcentaje
1	F	71.443	56,07 %
2	M	55.970	43,93 %
3	(Missing)	0	0,00 %
4	Total	12.7413	100,00 %

Nº	V47_TipoVivienda	Frecuencia	Porcentaje
1	Agricola Cultivo Propio Perenne	12	0,01 %
2	ARRENDADA	9.464	7,43 %
3	PRESTADA	197	0,15 %
4	PROPIA HIPOTECADA	388	0,30 %
5	PROPIA NO HIPOTECADA	44.049	34,57 %
6	VIVE CON FAMILIARES	72.756	57,10 %
7	(Missing)	547	0,43 %
8	Total	127.413	99,99 %

Nº	V48_EstadoCivil	Frecuencia	Porcentaje
1	CASADO	11.657	9,15 %
2	CONVIVIENTE	10.371	8,14 %
3	DIVORCIADO	15.830	12,42 %
4	SOLTERO	87.710	68,84 %
5	VIUDO	1.843	1,45 %
6	(Missing)	2	0,00 %
7	Total	127.413	100,00 %

Nº	V49_NivelEstudios	Frecuencia	Porcentaje
1	FORMACION INTERMEDIA	12.372	9,71 %
2	POSTGRADO	50	0,04 %
3	PRIMARIA	12.533	9,84 %
4	SECUNDARIA	96.573	75,8 %
5	SIN ESTUDIOS	1.362	1,07 %
6	UNIVERSIDAD	4.523	3,55 %
7	(Missing)	0	0,00 %
8	Total	127.413	100,00 %

Nº	V50_Sector	Frecuencia	Porcentaje
1	RUR	5.938	4,66 %
2	URB	12.1391	95,27 %
3	(Missing)	84	0,07 %
4	Total	127.413	100 %

Nº	V51_TieneTelefono	Frecuencia	Porcentaje
1	0	24	0,02 %
2	1	127.389	99,98 %
3	(Missing)	0	0,00 %
4	Total	127.413	100,00 %

Nº	V53_SectorEconomico	Frecuencia	Porcentaje
1	AGRICOLA	447	0,35 %
2	DEPENDENCIA	11.783	9,25 %
3	NEGOCIANTE	54.306	42,62 %
4	OFICIO	6.590	5,17 %
5	PERSONAL	1.200	0,94 %
6	PRODUCCION	17.257	13,54 %
7	SERVICIOS	35.828	28,12 %
8	(Missing)	2	0,00 %
9	Total	127.413	99,99 %

Nº	V56_Departamento_Dom	Frecuencia	Porcentaje
1	ANCASH	4	0,00 %
2	CAJAMARCA	2	0,00 %
3	CALLAO	6.154	4,83 %
4	HUANCAVELICA	46	0,04 %
5	HUANUCO	1	0,00 %
6	ICA	18.013	14,14 %
7	JUNIN	10.213	8,02 %
8	LAMBAYEQUE	8	0,01 %
9	LIMA	92.943	72,95 %
10	PASCO	26	0,02 %
11	TACNA	3	0,00 %
12	(Missing)	0	0,00 %
13	Total	127.413	100,01 %

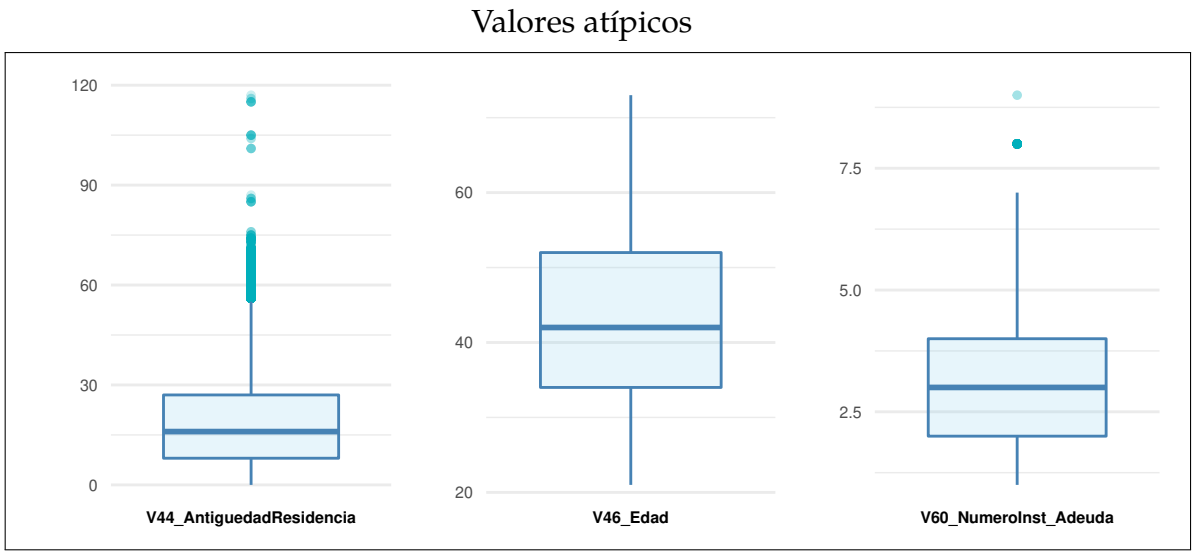
Nº	V59_Auto	Frecuencia	Porcentaje
1	0	105.313	82,65 %
2	1	22.100	17,35 %
3	(Missing)	0	0,00 %
4	Total	127.413	100,00 %

Nº	V63_TieneCreditoHipotecario	Frecuencia	Porcentaje
1	0	12.6923	99,62 %
2	1	490	0,38 %
3	(Missing)	0	0,00 %
4	Total	127.413	100,00 %

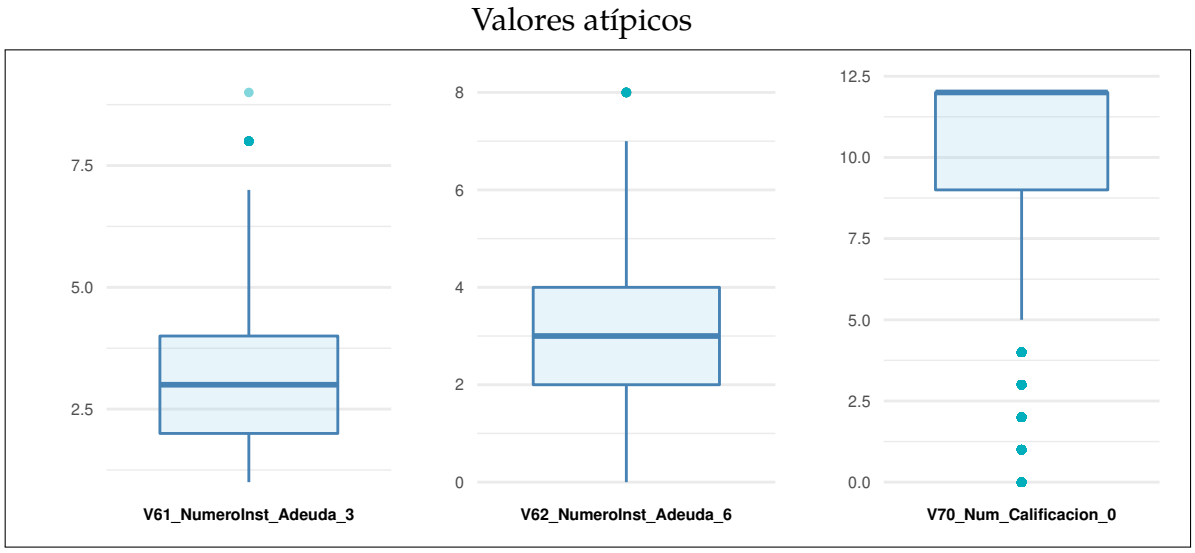
Nº	V69_TiempoHistorialCrediticioSF_6	Frecuencia	Porcentaje
1	5	41	0,03 %
2	6	127.372	99,97 %
3	(Missing)	0	0,00 %
4	Total	127.413	100,00 %

Nº	V100_TieneTarjetaCredito	Frecuencia	Porcentaje
1	0	67.780	53,2 %
2	1	59.633	46,8 %
3	(Missing)	0	0 %
4	Total	127.413	100 %

ANEXO 3: Diagramas de cajas y bigotes variables numéricas



Fuente: Elaboración propia.



Fuente: Elaboración propia.

ANEXO 4: Árboles de decisión

Figura 1
V17_CarteraRiesgo_Q

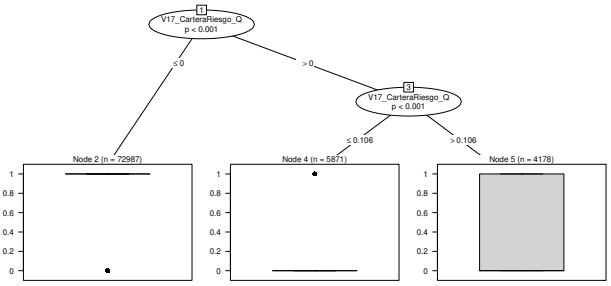


Figura 2
V18_SaldoMMora_Mont

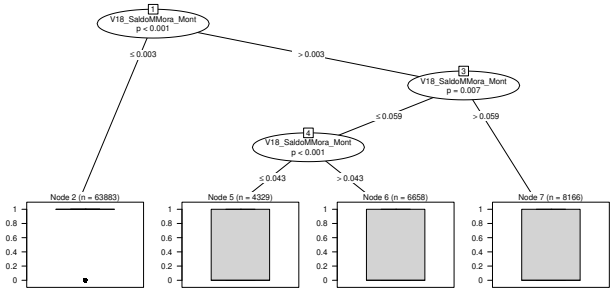


Figura 3
V25_Porc_CuotasPag

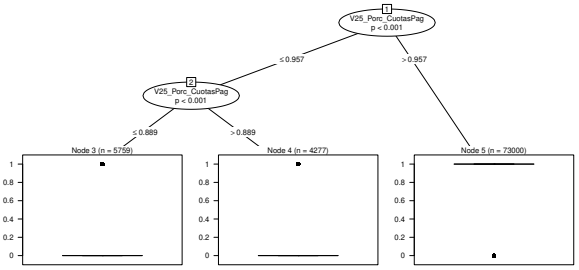


Figura 4
V26_Porc_CuotasVenc

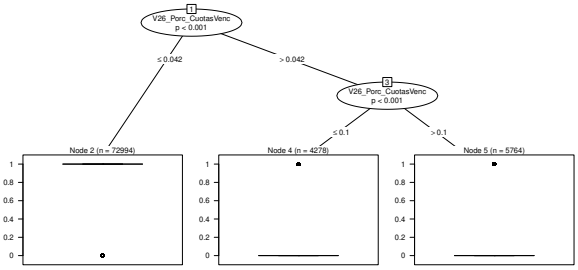


Figura 5
V29_PromAtraso_DC

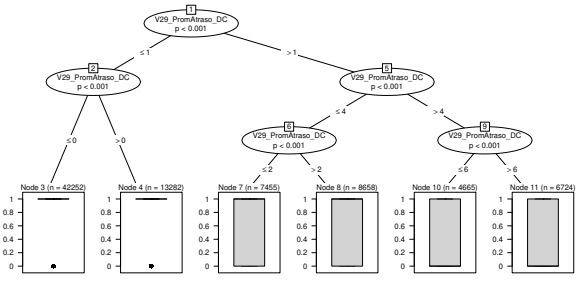
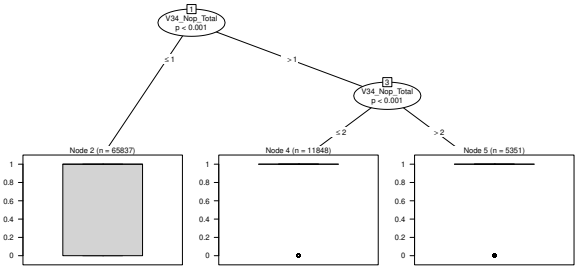


Figura 6
V34_Nop_Total



Fuente: Elaboración propia.

Figura 7

$V34\_Nop_Total\_v2_M8$

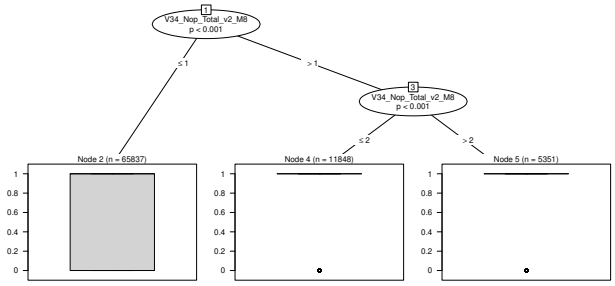


Figura 8

$V35\_Nop_Retanqueo\_Total$

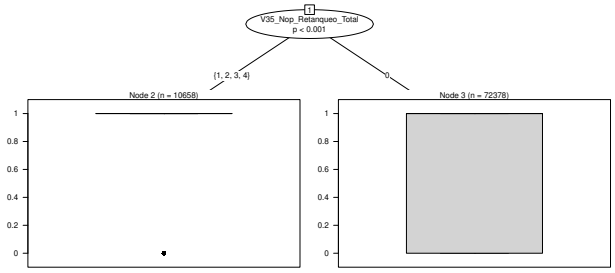


Figura 9

$V39\_CargasFamiliares$

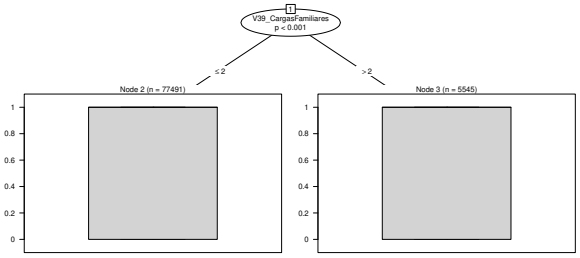


Figura 10

$V47\_TipoVivienda$

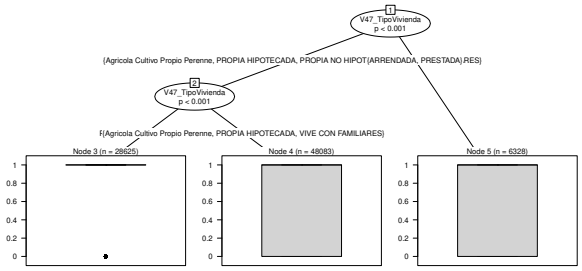


Figura 11

$V48\_EstadoCivil$

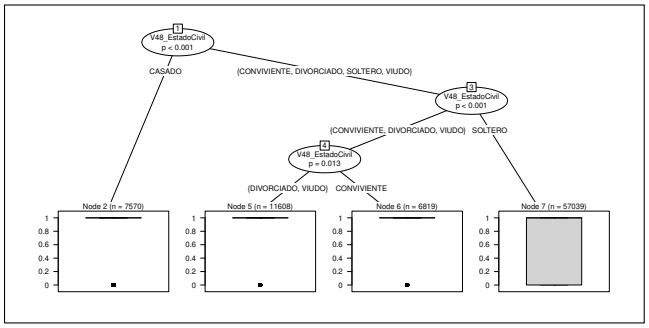


Figura 12

$V49\_NivelEstudios$

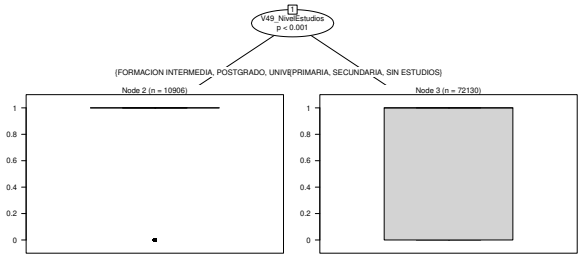


Figura 13

$V53\_SectorEconomico$

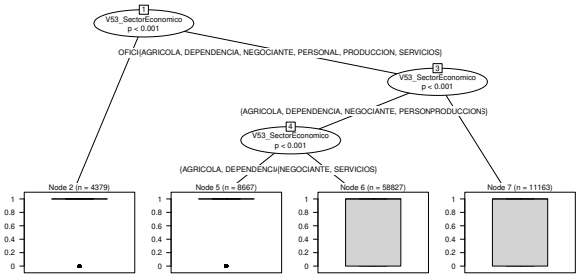
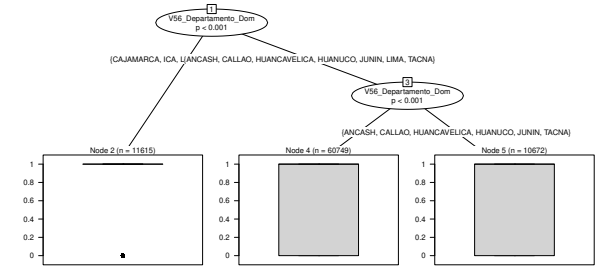


Figura 14

$V56\_Departamento\_Dom$



Fuente: Elaboración propia.

Figura 15
V57_Antiguedad_Laboral

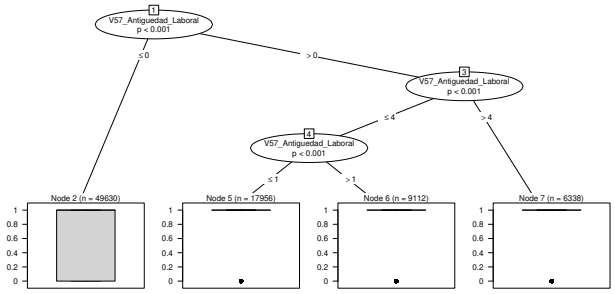


Figura 16
V65_TiempoHistorialCrediticioSF_36

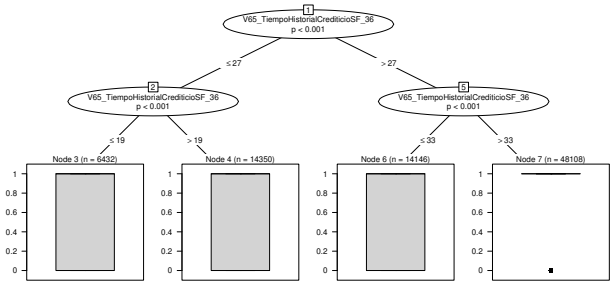


Figura 17
V66_TiempoHistorialCrediticioSF_24

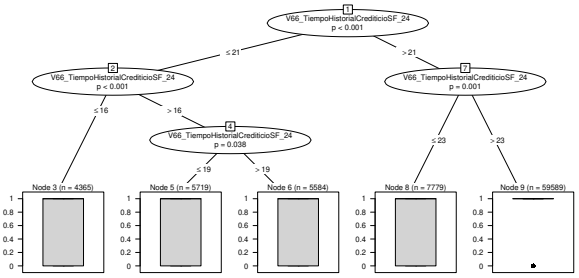


Figura 18
V67_TiempoHistorialCrediticioSF_18

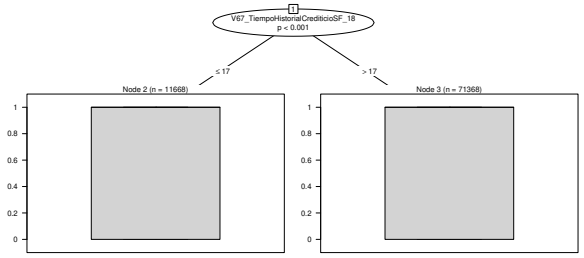


Figura 19
V71_Num_Calificacion_1

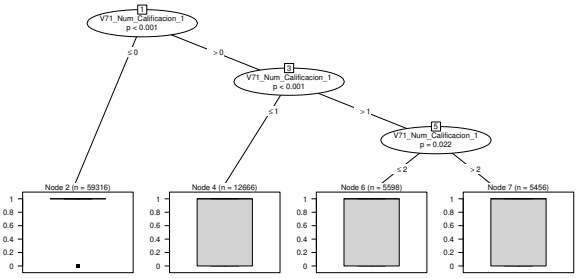


Figura 20
V72_Num_Calificacion_234

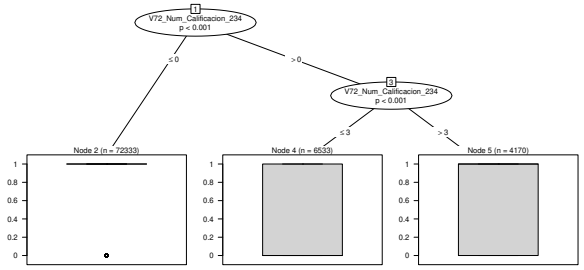


Figura 21
V73_PeorCalificacionCorte_Comp

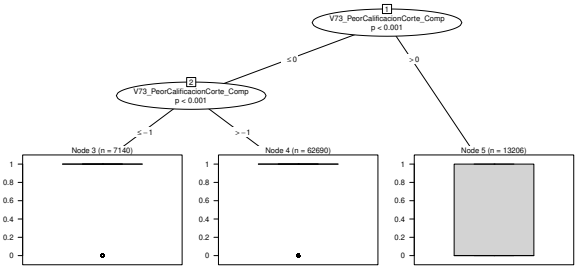
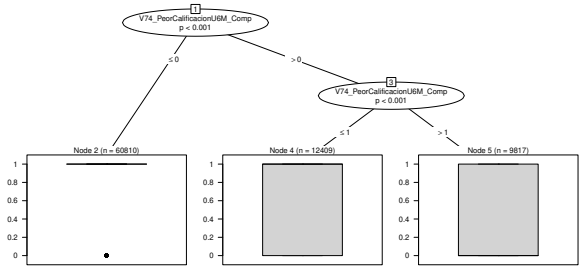


Figura 22
V74_PeorCalificacionU6M_Comp



Fuente: Elaboración propia.

Figura 23
V75_PeorCalificacionU12M_Comp

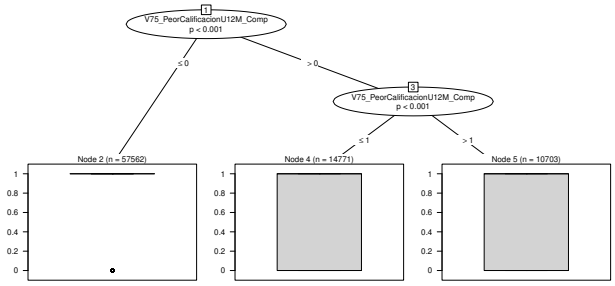


Figura 24
V76_PeorCalificacionU18M_Comp

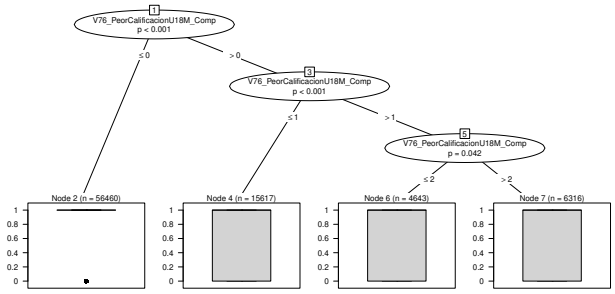


Figura 25
V77_CarteraRiesgoPond_U6M_Comp

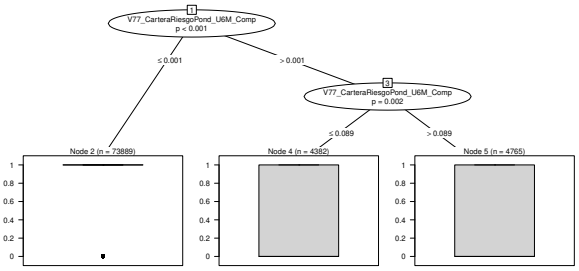


Figura 26
V78_CarteraRiesgo_comp

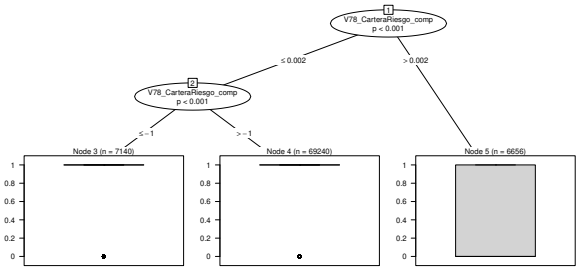


Figura 27
V79_MoraPonderada_comp

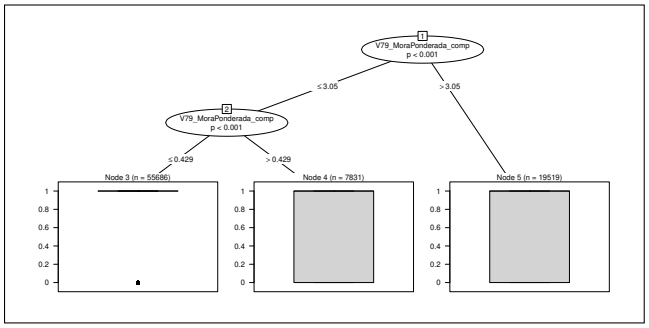


Figura 28
V82_Saldo_MB_Comp

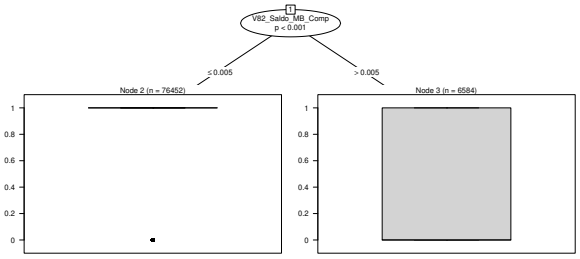


Figura 29
V85_VariacionDeudaConsumo_U3M

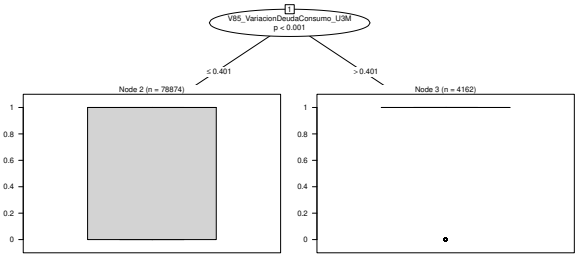
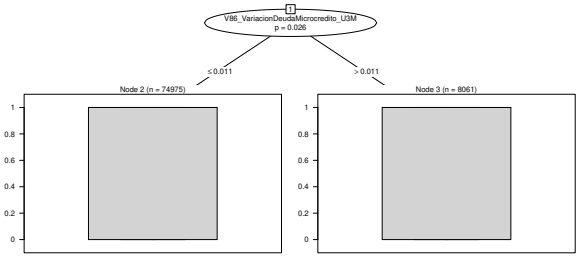


Figura 30
V86_VariacionDeudaMicrocredito_U3M



Fuente: Elaboración propia.

Figura 31

$V87\_VariacionDeudaTarjeta\_U3M$

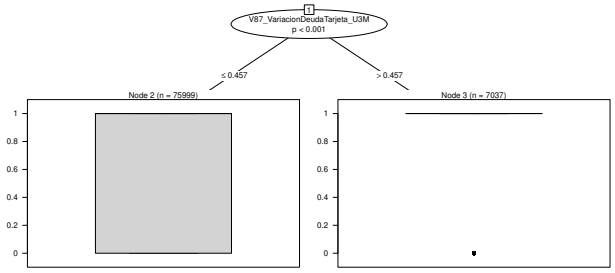


Figura 32

$V88\_VariacionDeudaTotal\_U3M$

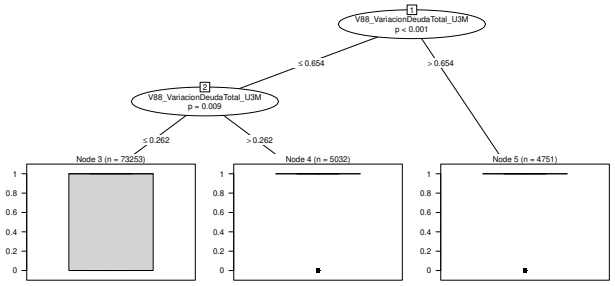


Figura 33

$V89\_VariacionDeudaConsumo\_U6M$

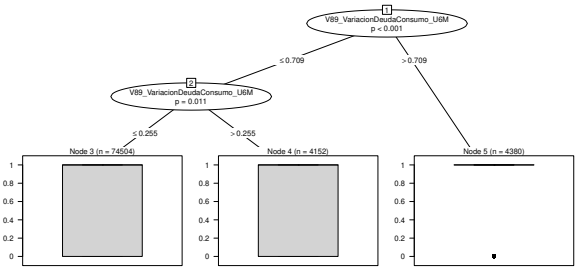
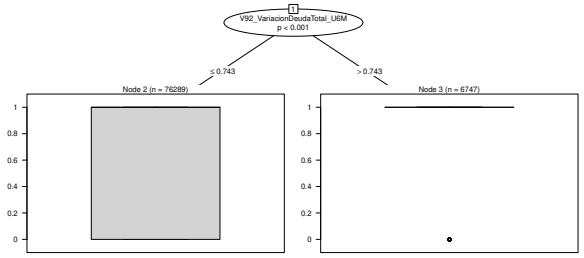


Figura 34

$V92\_VariacionDeudaTotal\_U6M$



Fuente: Elaboración propia.

ANEXO 5: Medidas de Asociación VI

Nº	Variable	VI
1	V29_PromAtraso_DC	1,248029
2	V17_CarteraRiesgo_Q	0,997835
3	V25_Porc_CuotasPag	0,984001
4	V26_Porc_CuotasVenc	0,983458
5	V18_SaldoMMora_Mont	0,768008
6	V79_MoraPonderada_comp	0,422633
7	V73_PeorCalificacionCorte_Comp	0,397356
8	V74_PeorCalificacionU6M_Comp	0,383076
9	V75_PeorCalificacionU12M_Comp	0,339025
10	V76_PeorCalificacionU18M_Comp	0,31697
11	V71_Num_Calificacion_1	0,293248
12	V78_CarteraRiesgo_comp	0,208485
13	V77_CarteraRiesgoPond_U6M_Comp	0,205314
14	V82_Saldo_MB_Comp	0,20399
15	V72_Num_Calificacion_234	0,191611
16	V53_SectorEconomico	0,059777
17	V57_Antiguedad_Laboral	0,050534
18	V65_TiempoHistorialCrediticioSF_36	0,030865
19	V48_EstadoCivil	0,030134
20	V34_Nop_Total	0,022423

Nº	Variable	VI
21	V34_Nop_Total_v2_M8	0,022423
22	V47_TipoVivienda	0,017543
23	V66_TiempoHistorialCrediticioSF_24	0,015043
24	V56_Departamento_Dom	0,012695
25	V49_NivelEstudios	0,009741
26	V35_Nop_Retanqueo_Total	0,009486
27	V88_VariacionDeudaTotal_U3M	0,00606
28	V39_CargasFamiliares	0,005663
29	V43_EsIndependiente	0,004149
30	V92_VariacionDeudaTotal_U6M	0,003648
31	V67_TiempoHistorialCrediticioSF_18	0,003476
32	V45_Genero	0,003334
33	V89_VariacionDeudaConsumo_U6M	0,003251
34	V85_VariacionDeudaConsumo_U3M	0,002455
35	V87_VariacionDeudaTarjeta_U3M	0,002346
36	V50_Sector	0,001168
37	V69_TiempoHistorialCrediticioSF_6	0,000823
38	V100_TieneTarjetaCredito	0,000638
39	V86_VariacionDeudaMicrocredito_U3M	0,00051
40	V63_TieneCreditoHipotecario	0,000501
41	V51_TieneTelefono	0,000389

ANEXO 6: Códido del algoritmo completo implementado en R

```
##############################################################################
####                              LIBRERIAS                               #####
##############################################################################
library(ggplot2)
library(data.table)
library(dplyr)
library(tidyverse)
library(DT)
library(lubridate)
library(writexl)
library(gridExtra)
library(caTools)
library(mgcv)
library(corrplot)
library(formattable)
library(stats)


##############################################################################
####                           BASE DE DATOS                              #####
##############################################################################
#### Base: Base de datos
Base <- fread(
  "BaseBrutoCC.csv",
```

```
  header = TRUE,
  sep = ";",
  dec = ".",
  na.strings =  "NULL",
  quote = '\t',
  stringsAsFactors = F,
  encoding = "UTF-8",
  data.table = FALSE,
  integer64 = "character"
)
Base$V1_FechaCorte <- as.Date(Base$V1_FechaCorte, format = "%d/%m/%Y")

#############################################################################
#################### ANÁLISIS DESCRIPTIVO UNIVARIADO ########################
#############################################################################
#### Análisis descriptivo - variables cuantitativas
AnalisisUniv_Cuant <- function(base){
  variable <- as.data.frame(dplyr::select_if(base, is.numeric))
  BaseCuantR <- NULL
  for (i in 1:ncol(variable)) {
    x <- select(variable, var=i)
    `Porc_NAs`<- round(sum(is.na(x$var))/length(x$var),4)*100
    `Porc_0s` <- round(sum(x$var=="0",na.rm=TRUE)/length(x$var),4)*100
    `Mínimo`  <- round(min(x$var,na.rm=TRUE),2)
    `Perc_25%`<- round(quantile(x$var,0.25,na.rm=TRUE),2)
    `Mediana` <- round(median(x$var,na.rm=TRUE),2)
    `Media`   <- round(mean(x$var,na.rm=TRUE),2)
    `Perc_75%`<- round(quantile(x$var,0.75,na.rm=TRUE),2)
    `Máximo`  <- round(max(x$var,na.rm=TRUE),2)
    `Desv_Est`<- round(sd(x$var,na.rm=TRUE),2)

    Aux <- tibble(`Porc_NAs`, `Porc_0s`, `Mínimo`, `Perc_25%`,
```

```
                  `Mediana`, `Media`, `Perc_75%`, `Máximo`, `Desv_Est`)
    BaseCuantR <- rbind(BaseCuantR, Aux)
  }
  BaseCuantR <- BaseCuantR %>%
  rownames_to_column("Variable") %>%
  rownames_to_column("Nº")
  BaseCuantR$Variable <- colnames(variable)
  return(BaseCuantR)
}

#### Análisis descriptivo - variables cualitativas
AnalisisUniv_Cual <- function(base) {
 base <- as.data.frame(base)
 NamesV <- names(base)
 StatM <- NULL
 for (i in 1:ncol(base)) {
   if (is.factor(base[, i]) | is.character(base[, i])) {
      StatV <- data.frame(table(base[, i], useNA = "always"))
      StatV$Var1 <- forcats::fct_explicit_na(StatV$Var1)
      if (is.null(StatM)==TRUE) {
        StatM <- data.frame(Variable = NamesV[i], StatV)
      } else {
        StatM <- rbind(StatM, data.frame(Variable = NamesV[i], StatV))
      }
    } else {
      next
    }
  }
  n <- nrow(base)
  StatM <- StatM %>%
    dplyr::rename(Variable = Variable, Categorias = Var1,
                  Frecuencia = Freq) %>%
```

```
    dplyr::group_by(Variable, Categorias, Frecuencia) %>%
    dplyr::summarise(Porcentaje = round(Frecuencia / n,4)*100) %>%
    as.data.frame()
  StatM <- data.frame(StatM)
  return(StatM)
}

##############################################################################
######################### DEPURACIÓN DE LOS DATOS ############################
##############################################################################
#### Imputación de variables
imputacion_variable <- function(variable, metodo = "ninguno") {
  if (any(IdNa <- is.na(variable))) {
    if (is.numeric(variable)) {
      if (metodo == "ninguno") {
        variable
      } else if (metodo == "media") {
        variable[IdNa] <- mean(variable, na.rm = TRUE)
      } else if (metodo == "mediana") {
        variable[IdNa] <- stats::median(variable, na.rm = TRUE)
      } else {
        stop("El método seleccionado no es correcto.")
      }
    } else if (is.factor(variable) | is.character(variable)) {
      if (metodo == "ninguno") {
        variable
      } else if (metodo == "moda") {
        variable[IdNa] <- levels(variable)[which.max(table(variable))]
      } else {
        stop("El método seleccionado no es correcto.")
      }
    } else {
```

```
      stop("La variable no es de tipo numeric ni de tipo factor.")
    }
  } else {
    stop("La variable no tiene valores perdidos.")
  }
  return(variable)
}

#### Imputación de variables Numéricas
psych::describe(Base$V57_Antiguedad_Laboral)
Base$V57_Antiguedad_Laboral <- imputacion_variable(
                    Base$V57_Antiguedad_Laboral, metodo = "mediana")

#### Imputación de variables categóricas
Base$V43_EsIndependiente <- imputacion_variable(
                    Base$V43_EsIndependiente, metodo = "moda")
Base$V47_TipoVivienda <- imputacion_variable(
                    Base$V47_TipoVivienda, metodo = "moda")
Base$V48_EstadoCivil <- imputacion_variable(
                    Base$V48_EstadoCivil, metodo = "moda")
Base$V50_Sector <- imputacion_variable(
                    Base$V50_Sector, metodo = "moda")
Base$V53_SectorEconomico <- imputacion_variable(
                    Base$V53_SectorEconomico, metodo = "moda")

#### Análisis de Consistencia
nombVar<-c("V25_Porc_CuotasPag",
           "V85_VariacionDeudaConsumo_U3M",
           "V86_VariacionDeudaMicrocredito_U3M",
           "V87_VariacionDeudaTarjeta_U3M",
           "V88_VariacionDeudaTotal_U3M",
           "V89_VariacionDeudaConsumo_U6M",
```

```
            "V90_VariacionDeudaMicrocredito_U6M",
            "V91_VariacionDeudaTarjeta_U6M",
            "V92_VariacionDeudaTotal_U6M"
)

Consistencia <- function(base,nomVar){
  for(variable in nombVar) {
    Aux <- select(base,NomVar=i)
    Aux$NomVar <- ifelse(Aux$NomVar<0,0,ifelse(Aux$NomVar>1,1, Aux$NomVar))
    base[,i] <- Aux
  }
  return(base)
}
summary(Base$V46_Edad)

#### Análisis de variables constantes
VarConstantes <- function(base){
  variables <- dplyr::select_if(dplyr::select(base, -c(1,2,3)),
   negate(is.character))
  variables <- dplyr::select_if(variables, negate(is.factor))
  Var <- NULL
  for (i in names(variables)) {
    Aux <- dplyr::select(variables, NomVar=i)
    if(min(Aux$NomVar, na.rm = T)==max(Aux$NomVar, na.rm = T) |
        any(prop.table(table(Aux$NomVar))>=0.97)){
      Var <- c(Var, i)
    }
  }
  return(Var)
}
Var <- VarConstantes(Base)
Base <- dplyr::select(Base, -Var)
```

```
##############################################################################
#################### TRATAMIENTO DE DATOS ATÍPICOS ###########################
##############################################################################
#### Método de Winsornización
winsorizing_method <- function(variable, removeNA = TRUE){
  percentil <- quantile(variable,
                        probs = c(.05, .95),
                        type=1, na.rm = removeNA)
  variable[variable<percentil[1]] <- percentil[1]
  variable[variable>percentil[2]] <- percentil[2]
  return(variable)
}

##############################################################################
######################### DESARROLLO - VALIDACION ############################
##############################################################################
#### Muestras: Desarrollo / Validacion (70%-30%)
##Base sin Indeterminados
Base1 <- dplyr::filter(Base, Variable_Dependiente != 2)

set.seed(123)
index <- sample(1:nrow(Base1),
                replace = FALSE,
                size = floor(0.7*nrow(Base1)))
bd_des <- Base1[index,]
bd_val <- Base1[-index,]

##############################################################################
################### CATEGORIZACIÓN Y RECATEGORIZACIÓN ########################
##############################################################################
#### Categorización de variables cuantitativas y cualitativas mediante
```

```r
#### árboles de decisión
categorizacion <- function (base, resp, n_per = 7, n_factor = 5) {
  nombres <- colnames(base)[colnames(base) != resp]
  BDDnew_cat <- data.frame(Variable = character(0),
                           NuevaVariable = character(0),
                           Formula = character(0),
                           Detalle = character(0))
  for (i in nombres) {
    if (is.numeric(base[, i])) {
      prove_percentil <- Hmisc::cut2(base[, i], g = n_per)
      if (length(levels(prove_percentil)) != n_per) {
        regla <- extraccion_reglas_arbol(base, resp, i)
        new_nom_var <- paste(i, "cat", sep = "_")
        BDDnew_cat <- rbind(BDDnew_cat,
                            data.frame(Variable = i,
                                       NuevaVariable = new_nom_var,
                                       Formula = regla,
                                       Detalle = "pretransformacion"))
      } else {
        next()
      }
    } else {
      if (length(levels(droplevels(as.factor(base[, i])))) >= n_factor) {
        regla <- extraccion_reglas_arbol(base,
                                         respuesta = resp,
                                         variable = i)
        new_nom_var <- paste(i, "cat", sep = "_")
        BDDnew_cat <- rbind(BDDnew_cat,
                            data.frame(Variable = i,
                                       NuevaVariable = new_nom_var,
                                       Formula = regla,
                                       Detalle = "pretransformacion"))
```

```
      } else {
        next()
      }
    }
  }
  BDDnew_cat$Formula <- as.character(BDDnew_cat$Formula)
  BDDnew_cat <- dplyr::filter(BDDnew_cat, Formula != "Sin_Arbol")
  return(BDDnew_cat)
}

#### Extracción de las reglas de decisión de un árbol
extraccion_reglas_arbol <- function(base, resp, variable, n_porc = 0.05) {
  Y <- as.character(resp)
  x <- as.character(variable)
  BDD <- as.data.frame(base[, c(Y, x)])
  individuos_nodos <- round(n_porc * nrow(BDD))
  formula_ctree <- formula(paste(Y, x, sep = " ~ "))
  numbers_only <- function(x){
    suppressWarnings(!is.na(as.numeric(as.character(x))))
    }
  try(
    if (is.numeric(BDD[, x])) {
      ct1 <- partykit::ctree(formula_ctree, data = BDD,
        control = partykit::ctree_control(minbucket = individuos_nodos))
      Nodo1 <- as.character(names(partykit:::.list.rules.party(ct1)))
      if (length(Nodo1) > 1) {
        Regla1 <- as.character(partykit:::.list.rules.party(ct1))
        sp1 <- strsplit(Regla1, split = " ")
        sp2 <- list()
        for (i in 1:length(sp1)) {
          aux <- sp1[[i]]
          sp2[[i]] <- aux[numbers_only(sp1[[i]])]
```

```
      }
      valores <- sp2
      valores_unicos <- sort(unique(as.numeric(unlist(valores))))
      new_variable <- cut(BDD[, x],
                          breaks = c(-Inf, valores_unicos, Inf),
                          ordered_result = TRUE, dig.lab = 5)
      tabla_categorias <- data.frame(Variable = x,
                            Categorias = sort(unique(new_variable)),
                            Valores = Regla1,
                            stringsAsFactors = FALSE)
      Regla2 <- NULL
      n <- length(Regla1) - 1
      for (i in 1:n) {
        Regla2[i] <- paste0("ifelse(", Regla1[i], ", ",
                            paste0("c(", "\"",
                            tabla_categorias$Categorias[i], "\"", ")"))
      }
      formula1 <- paste(Regla2, collapse = ", ")
      cierre_formula <- paste0(rep(")", n), collapse = "")
      formula2 <- paste(formula1,
                        paste0("c(", "\"",
                        tabla_categorias$Categorias[length(Regla1)],
                        "\"", ")"),  sep = ", ")
      formula_final <- paste0(formula2, cierre_formula)
    } else {
      formula_final <- NULL
    }
  } else if (is.factor(BDD[, x])) {
    BDD[, x] <- droplevels(BDD[, x])
    ct1 <- party::ctree(formula_ctree, data = BDD,
          controls = party::ctree_control(minbucket = individuos_nodos))
    aux <- data.frame(var = BDD[, x], Nodo = party::where(ct1))
```

```
    Nodos <- sort(unique(aux$Nodo))
    if (length(Nodos) > 1) {
      grupos <- list()
      for (i in 1:length(Nodos)) {
        aux_0 <- dplyr::filter(aux, Nodo == Nodos[i])
        grupos[[i]] <- sort(unique(as.character(aux_0[, "var"])))
      }
      Category <- paste("\"", paste("Grupo_", 1:length(Nodos),
                                    sep = ""), "\"", sep = "")
      tabla_categorias <- data.frame(Variable = x,
                                     Categorias = Category[1],
                                     Valores = paste(grupos[[1]],
                                                     collapse = ", "),
                                     stringsAsFactors = FALSE)
      for (i in 2:length(Category)) {
        tabla_categorias <- dplyr::bind_rows(tabla_categorias,
                                     data.frame(Variable = x,
                                         Categorias = Category[i],
                                         Valores = paste(grupos[[i]],
                                                         collapse = ", "),
                                         stringsAsFactors = FALSE))
      }
      vector_grupos <- list()
      for (j in 1:length(grupos)) {
        vector_grupos[[j]] <- paste("c(", (paste0(paste("\"",
                                    grupos[[j]],"\"", sep = ""),
                                    collapse = ", ")), ")",
                                    sep = "")
      }
      n <- length(grupos) - 1
      Regla2 <- NULL
      for (i in 1:n) {
```

```
          Regla2[i] <- paste0(paste("ifelse(", x, " %in% ",
                              vector_grupos[[i]], sep = ""), ", ",
                              Category[i])
        }
        formula1 <- paste(Regla2, collapse = ", ")
        cierre <- paste0(rep(")", n), collapse = "")
        ReglaFinal <- paste(formula1, Category[length(grupos)],
                              sep = ", ")
        formula_final <- paste0(ReglaFinal, cierre)
      } else {
        formula_final <- NULL
      }
    }, silent = TRUE
  )
  if (is_empty(formula_final)) {
    return("Sin_Arbol")
  } else {
    return(Regla = formula_final)
  }
}

##############################################################################
############### SELECCIÓN DE VARIABLES EXPLICATIVAS (KS -IV) #############
##############################################################################
#### Estadístico Kolmogórov-Smirnov
# resp: variable dependiente.
# base: variables explicativas.
KS_test <- function(resp, base) {
  Y <- base[, resp]
  variable <- as.data.frame(select_if(base[, setdiff(names(base),
                                                     resp)],is.numeric))
  Aux <- NULL
```

```
  for (i in 1:ncol(variable)) {
    m <- data.frame(Y, variable[,i])
    m1 <- filter(m, Y == "1")
    m2 <- filter(m, Y == "0")
    ks <- suppressWarnings(
      stats::ks.test(
        m1[, 2],
        m2[, 2],
        alternative = "two.sided",
        exact = FALSE
      )
    )
    Aux1 <- data.frame(Valor_KS =round(as.numeric(ks$statistic), 4))
    Aux <- rbind(Aux, Aux1)
  }
  Tabla_KS <- tibble::rownames_to_column(Aux, "Variable")
  Tabla_KS$Variable  <- names(variable)
  Tabla_KS <- dplyr::arrange(Tabla_KS, desc(Valor_KS))
  return(Tabla_KS)
}

#### Calcula el estadístico de Information-Value
# resp: variable dependiente.
# base: variables explicativas.
Information_Value <- function(resp, base){
  Y <- base[, resp]
  variable <- as.data.frame(select_if(base[, setdiff(names(base),
                                                     resp)], is.factor))
  IV <- NULL
  for (i in 1:ncol(variable)) {
    Frec <- table(Y,variable[,i])
    Frec1 <- Frec[1,]
```

```
    Frec2 <- Frec[2,]
    Aux1 <- ifelse(Frec1/sum(Frec1)==0, 0.0001,
                   ifelse(Frec1/sum(Frec1)==1, 0.999,
                          Frec1/sum(Frec1)))
    Aux2  <- ifelse(Frec2/sum(Frec2)==0, 0.0001,
                    ifelse(Frec2/sum(Frec2)==1, 0.999,
                           Frec2/sum(Frec2)))
    Woe <- log(Aux2/Aux1)
    Woe <- ifelse(Woe==-Inf, 0, Woe)
    IV1 <- data.frame(Valor_IV = sum((Frec2/sum(Frec2) -
                                        Frec1/sum(Frec1))*Woe))
    IV <- rbind(IV, IV1)
  }
  Tabla_IV <- tibble::rownames_to_column(IV, "Variable")
  Tabla_IV$Variable  <- names(variable)
  Tabla_IV <- dplyr::arrange(Tabla_IV, desc(Valor_IV))
  return(Tabla_IV)
}

################################################################################
##### MODELOS ADITIVOS GENERALIZADOS (Splines Cúbicos de Regresión) ####
############################## MODELO GAM-2 ####################################
################################################################################
#### Modelo logístico aditivo generalizado óptimo
log_mod_GAM2 <- gam(Variable_Dependiente ~
                    ##### Numéricas
                    V16_Amortizacion +
                    s(V31_AtrasoMax_U6_AC, bs = 'cr', k = 12) +
                    V38_AntiguedadCliente +
                    V46_Edad +
                    V60_NumeroInst_Adeuda +
                    s(V81_Amortizacion_comp, bs = 'cr', k =6) +
```

```
                V90_VariacionDeudaMicrocredito_U6M +
                V91_VariacionDeudaTarjeta_U6M +
                ##### Categóricas
                V17_CarteraRiesgo_Q +
                V18_SaldoMMora_Mont +
                V29_PromAtraso_DC +
                V79_MoraPonderada_comp +
                V82_Saldo_MB_Comp,
                data = bd_train,
                family = binomial,
                method = "REML",
                select = TRUE
              )

##############################################################################
######################### EVALUACIÓN ESTADÍSTICA #############################
##############################################################################
#### Medidas de discriminación (AUC, KS, GINI)
prediccion <- predict(log_mod_GAM2, type = "response", newdata = bd_test)
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
MLmetrics::AUC(prediccion, respuesta)
MLmetrics::KS_Stat(prediccion, respuesta)/100
MLmetrics::Gini(prediccion, respuesta)

### Matriz de confusión
prediccion <- as.factor(ifelse(predict(log_mod_GAM2,
                        type = "response",
                        newdata = bd_test) < corte_optimo, 0, 1))
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
caret::confusionMatrix(prediccion, respuesta, positive = '1')

### Tabla de desempeño
```

```
Tabla_Performance <- function(prediccion, respuesta, n = 10){
  Performance <- data.frame(prediccion, respuesta) %>%
    dplyr::mutate(respuesta = as.numeric(as.character(respuesta)),
                  Deciles = factor(Hmisc::cut2(prediccion, g = n,
                                               digits = 3))) %>%
    dplyr::group_by(Deciles) %>%
    dplyr::summarise(N = n(), Porc_N = round(n()/length(obj),4)*100) %>%
    dplyr::mutate(Acum_N=cumsum(Porc_N))
  Aux <- as.data.frame.matrix(table(Hmisc::cut2(prediccion,
                                                g = n,
                                                digits = 3),
                                    respuesta)) %>%
    rownames_to_column("Deciles")
  Performance <- data.frame(Performance, Buenos=Aux[,3],
                            row.names=NULL) %>%
    dplyr::mutate(Porc_Buenos = round(Buenos/sum(Buenos),3)*100) %>%
    dplyr::mutate(Acum_Buenos = cumsum(Porc_Buenos))
  Performance <- data.frame(Performance, Malos=Aux[,2],
                            row.names=NULL) %>%
    dplyr::mutate(Porc_Malos = round(Malos/sum(Malos),3)*100) %>%
    dplyr::mutate(Acum_Malos = cumsum(Porc_Malos)) %>%
    dplyr::mutate(Razon_Buenos = round(Buenos/N,5)*100) %>%
    dplyr::mutate(Razon_Malos = round(Malos/N,3)*100)
  return(Performance)
}

#########################################################################
############### MODELO DE REGRESIÓN LOGÍSTICA (MÉTODO  1) ###############
############################## MODELO RL-M1 ##############################
#########################################################################
#### Se emplean la mismas muestras de modelamiento y validación
#### Modelo logístico bajo el método 1 (Entran las variables del modelo
```

```
#### GAM-2)
log_mod_RLM1 <- glm(Variable_Dependiente ~
                    ##### Numéricas
                    V16_Amortizacion +
                    V31_AtrasoMax_U6_AC +
                    I(V31_AtrasoMax_U6_AC^2) +
                    V38_AntiguedadCliente +
                    V46_Edad +
                    V60_NumeroInst_Adeuda +
                    V81_Amortizacion_comp +
                    I(V81_Amortizacion_comp^2) +
                    V90_VariacionDeudaMicrocredito_U6M +
                    V91_VariacionDeudaTarjeta_U6M +
                    ##### Categóricas
                    V17_CarteraRiesgo_Q +
                    V18_SaldoMMora_Mont +
                    V29_PromAtraso_DC +
                    V79_MoraPonderada_comp +
                    V82_Saldo_MB_Comp,
                    data = bd_train,
                    family = binomial(link="logit")
                    )

#### Medidas de discriminación (AUC, KS, GINI)
prediccion <- predict(log_mod_RLM1, type = "response", newdata = bd_test)
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
MLmetrics::AUC(prediccion, respuesta)
MLmetrics::KS_Stat(prediccion, respuesta)/100
MLmetrics::Gini(prediccion, respuesta)

### Matriz de confusión
prediccion <- as.factor(ifelse(predict(log_mod_RLM1,
```

```
                                type = "response",
                                newdata = bd_test) < corte_optimo, 0, 1))
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
caret::confusionMatrix(prediccion, respuesta, positive = '1')

### Tabla de desempeño
prediccion <- as.factor(ifelse(predict(log_mod_RLM1,
                                type = "response",
                                newdata = bd_test) < corte_optimo, 0, 1))
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
tabla_desempeño <- Tabla_Performance(prediccion, respuesta, n = 10)

##############################################################################
############### MODELO DE REGRESIÓN LOGÍSTICA (MÉTODO  2) ###############
############################## MODELO RL-M2 ##################################
##############################################################################
#### Se emplean la mismas muestras de modelamiento y validación
#### Modelo logístico bajo el método 2 (Regresión logística tradicional)
log_mod_RLM2 <- glm(Variable_Dependiente ~
                    ##### Numéricas
                    V16_Amortizacion +
                    V31_AtrasoMax_U6_AC +
                    V38_AntiguedadCliente +
                    V46_Edad +
                    V60_NumeroInst_Adeuda +
                    V68_TiempoHistorialCrediticioSF_12 +
                    V81_Amortizacion_comp +
                    V90_VariacionDeudaMicrocredito_U6M +
                    V91_VariacionDeudaTarjeta_U6M +
                    ##### Categóricas
                    V18_SaldoMMora_Mont +
                    V25_Porc_CuotasPag +
```

```
                    V29_PromAtraso_DC +
                    V79_MoraPonderada_comp +
                    V82_Saldo_MB_Comp,
                    data = bd_train,
                    family = binomial(link="logit")
                  )

#### Medidas de discriminación (AUC, KS, GINI)
prediccion <- predict(log_mod_RLM2, type = "response", newdata = bd_test)
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
MLmetrics::AUC(prediccion, respuesta)
MLmetrics::KS_Stat(prediccion, respuesta)/100
MLmetrics::Gini(prediccion, respuesta)

### Matriz de confusión
prediccion <- as.factor(ifelse(predict(log_mod_RLM2,
                               type = "response",
                               newdata = bd_test) < corte_optimo, 0, 1))
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
caret::confusionMatrix(prediccion, respuesta, positive = '1')

### Tabla de desempeño
prediccion <- as.factor(ifelse(predict(log_mod_RLM2,
                               type = "response",
                               newdata = bd_test) < corte_optimo, 0, 1))
respuesta <- bd_test %>% pull("Variable_Dependiente") %>% as.factor()
tabla_desempeño <- Tabla_Performance(prediccion, respuesta, n = 10)
```