ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELO DE ACTIVACIÓN DE TARJETAS DE CRÉDITO EN EL MERCADO CREDITICIO ECUATORIANO A TRAVÉS DE UNA METODOLOGÍA ANALÍTICA Y AUTOMATIZADA EN R.

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

ALEX EFRÉN PÉREZ TATAMUÉS

alx_perez90@hotmail.com

DIRECTOR: LUIS ALCIDES HORNA HUARACA, PhD.

luis.horna@epn.edu.ec

Quito, Diciembre 2014

DECLARACIÓN

Yo, Alex Efrén Pérez Tatamués, 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.

ALEX EFRÉN PÉREZ TATAMUÉS

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Alex Efrén Pérez Tatamués, bajo mi supervisión.

Luis Horna, PhD.
DIRECTOR DE PROYECTO

AGRADECIMIENTOS

Gracias a mis padres Luis Pérez y Zoila Tatamués, por su inmenso sacrificio y entrega, con el firme propósito de que pueda alcanzar esta meta tan importante en mi vida.

Gracias al Dr. Luis Horna por su acertada dirección, y valioso apoyo durante el desarrollo de este trabajo.

Gracias a los docentes de la Facultad de Ciencias, quienes con su enseñanza me permitieron culminar mi carrera universitaria.

Gracias a Diego, Julita, Carito, Jose, por sus importantes sugerencias y aportes durante el desarrollo del presente estudio.

DEDICATORIA

A Dios por la bendición de permitirme alcanzar este logro anhelado.

A mis padres Luis y Zoila, por su amor y sacrificio, quienes con su enseñanza y ejemplo han hecho de mi la persona que soy hoy en día.

A mi única hermana Edith, quien me ha apoyado en toda situación, en nuestra vida lejos de casa.

A mi familia y amigos, quienes con su cariño y palabras de apoyo me motivaron a terminar mi carrera universitaria.

Índice general

Índice de figuras

Índice de cuadros

Resumen

En el presente proyecto se describe una metodología estadística basada en medidas de divergencia, medidas de asociación, árboles de decisión y regresión logística; la cual es muy empleada en la construcción de modelos de activación de tarjetas de crédito. Estos modelos toman en consideración el hábito de consumo actual e histórico que tiene un sujeto, con el fin de predecir la probabilidad de que dicho sujeto realice al menos un consumo con una nueva tarjeta, en una ventana de tiempo determinada posterior a la fecha de apertura de la misma. Adicionalmente, se implementa un algoritmo en el software estadístico R, el cual se encarga de realizar de manera automática cada uno de los pasos de la metodología descrita.

Palabras claves: Medidas de divergencia, medidas de asociación, árboles de decisión, regresión logística, programación en R.

Abstract

The purpose of this project is describe a statistic methodology based on divergent measures, association measures, decision trees and logistic regression, that is frequently used in the construction of credit card activation models. These models considerate the current and historic consumption that have someone in order to predict the probability that this person make at least one usage with a new credit card during a determinated time after the opening date of it. Additionally, it will be implemented an algorithm in the statistical software R, this model performs the described methodology automatically step by step.

Key words: Divergent measures, association measures, decision trees, logistic regression, R programming.

Capítulo 1

Introducción.

En los últimos años se ha evidenciado un incremento notable en el endeudamiento en tarjetas de crédito que registra la población crediticia ecuatoriana. La competencia entre las principales instituciones emisoras de este servicio ha provocado que dichas instituciones desarrollen una serie de estrategias destinadas a adquirir nuevos y a mantener sus clientes actuales. Una de las estrategias de adquisición y retención de clientes más empleada consiste en asignar un cupo superior al ofertado por sus principales competidores; a pesar de ser una estrategia muy eficiente para la entidad, para el cliente representa un riesgo, debido a que el disponer de un cupo muy elevado puede ocasionar que se haga uso del servicio hasta llegar al punto de que la deuda adquirida supere al patrimonio disponible (sobreendeudamiento).

En la Figura 1.1 se presenta la evolución de la deuda y el número de tarjetas de crédito de la población crediticia ecuatoriana en el periodo comprendido entre octubre 2011 y octubre 2014[1]. Si analizamos la Figura 1.1a, tenemos que la deuda en tarjetas de crédito presenta un incremento notable en los últimos tres años; un comportamiento similar ocurre al analizar la evolución del número de tarjetas de crédito en la Figura 1.1b.

En el periodo comprendido entre junio 2012 y marzo 2013 (Figura 1.2) se tiene un comportamiento en particular. En la Figura 1.2 observamos que a hasta antes de

[1]La información utilizada se encuentra disponible en la página web de la Superintendencia de Bancos y Seguros `http://www.sbs.gob.ec`

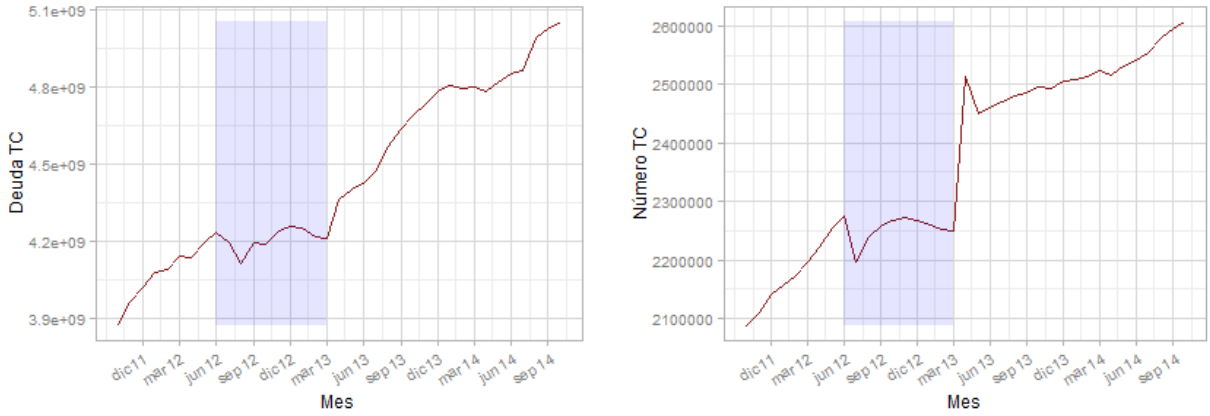

Figura 1.1: Evolución de Deuda y número de TC en el periodo oct-11 a oct-14

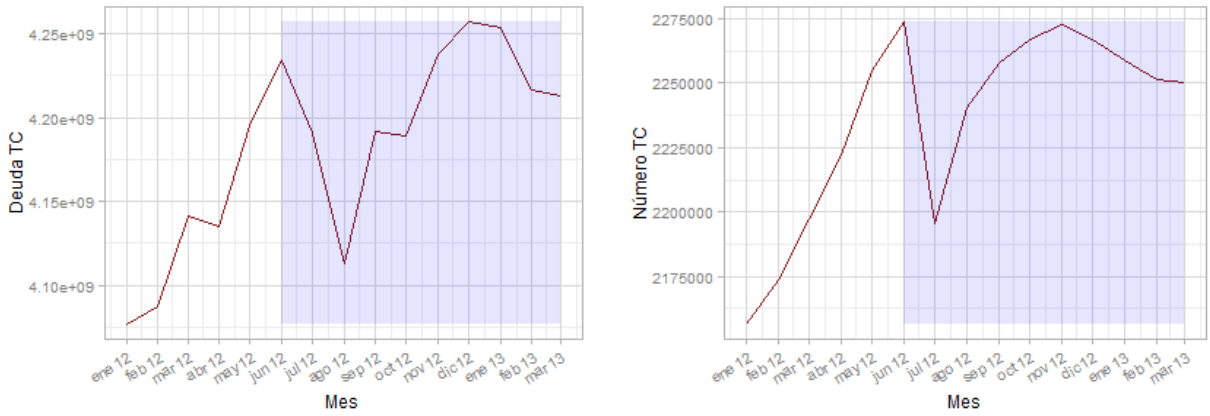

Figura 1.2: Evolución de Deuda y número de TC en el periodo jun-12 a mar-13

junio del 2012 tanto la deuda como el número de tarjetas presentaban un crecimiento acelerado, al evidenciar este crecimiento pronunciado y la posible presencia de sobreendeudamiento en el mercado crediticio ecuatoriano, el ente de control decide intervenir y tomar las acciones correspondientes con el propósito de **minimizar el riesgo de sobreendeudamiento de los ecuatorianos** y es así que en junio de 2012 la Junta Bancaria del Ecuador[2], considerando la necesidad de determinar límites de crédito en las operaciones de consumo incluyendo las operaciones de tarjeta de crédito (TC) y la necesidad de establecer patrimonio técnico sobre los créditos de consumo aprobados no desembolsados incluyendo los cupos de TC no utilizados,

[2]Organismo bajo el control de la Superintendencia de Bancos y Seguros, encargada de formular políticas de control y supervisión, regulaciones y resoluciones de aspectos que no estén claros en la Ley General de Instituciones del Sistema Financiero y que necesiten ser modificados.

mediante la resolución $JB - 2012 - 2217$[3] resuelve efectuar varias modificaciones en el **libro I** "Normas generales para la aplicación de la Ley General de Instituciones del Sistema Financiero[4]", específicamente en el **título V** "Del patrimonio técnico".

El patrimonio técnico requerido es el valor monetario que tiene como objetivo mantener constantemente la solvencia de la institución, su cálculo está asociado con el total de la suma de activos y contingentes ponderados por riesgo, generalmente este valor es fijado por el ente regulador.

La provisión, en cambio, es un valor monetario necesario para disminuir los efectos eventuales de pérdidas provocadas por incobrabilidad, está asociada al estado de resultados, generalmente su cálculo se basa en la calificación de riesgo o en la cuantificación del mismo mediante el uso de una metodología matemática.

La motivación del presente estudio se basa en una modificación, realizada mediante la resolución, en la forma de cálculo del patrimonio técnico requerido. La modificación que abordamos en este trabajo es la que concierne al cambio en el factor de ponderación de la cuenta **6404 Créditos aprobados no desembolsados** para efectos del cálculo del patrimonio técnico requerido, la modificación realizada se detalla a continuación:

Actualmente el nivel mínimo de patrimonio técnico requerido exigido por la SBS a las instituciones financieras es del $9\,\%$ de la suma total de los activos y contingentes ponderados por riesgo. Dependiendo del tipo de activo y contingente el factor de ponderación varía tal como muestra la Tabla 1.1, la misma que se elaboró en base a la información que presenta el **libro I** de la Ley General de Instituciones del Sistema Financiero mencionado anteriormente.
Considerando la Tabla 1.1, el patrimonio técnico requerido (PTR) se obtiene mediante la expresión:

[3]La resolución en su formato original se puede obtener en la página web de la SBS. http://www.sbs.gob.ec/medios/PORTALDOCS/downloads/normativa/2012/resol_JB-2012-2217.pdf

[4]El documento en su formato original se puede obtener en la página web oficial de la SBS. http://www.sbs.gob.ec/practg/sbs_index?vp_tip=12

Activos y contingentes	Factor Ponderación	Monto Activo
Fondos disponibles, IVA, Créditos aprobados no desembolsados (Reformado con resolución No. $JB-2012-2217$ de 22 de junio del 2012)	0,00	m1
Títulos crediticios emitidos o garantizados por el Estado o el Banco Central del Ecuador	0,10	m2
Títulos crediticios emitidos o garantizados por otras instituciones financieras del sector público	0,20	m3
Avales, fianzas	0,40	m4
Préstamos para la vivienda respaldados por hipoteca, arrendamiento inmobiliario e inversión en cédulas hipotecarias	0,50	m5
Colocaciones en préstamos o títulos crediticios y demás activos e inversiones físicas y financieras	1,00	m6

Tabla 1.1: Factor de ponderación para el cálculo del Patrimonio técnico

$$PTR = 9\,\%\,(0,0\,m1 + 0,10\,m2 + 0,20\,m3 + 0,40\,m4 + 0,50\,m5 + 1\,m6) \tag{1.1}$$

Hasta antes de junio del 2012, las instituciones financieras consideraban como factor de ponderación de la cuenta 6404 Créditos aprobados no desembolsados, el valor de cero $(0,0)$, es decir el monto de esta cuenta no se consideraba en el cálculo del patrimonio técnico requerido.

Luego de conocerse la resolución $JB-2012-2217$ de la Junta Bancaria en junio del 2012, mediante la cual se resolvió efectuar los siguientes cambios:

- Excluir de la cuenta 6404 Créditos aprobados no desembolsados, los valores correspondientes a créditos de consumo.
- Añadir la subcuenta 640401 Créditos aprobados no desembolsados - Cartera de créditos de consumo.
- Establecer que las instituciones emisoras de tarjetas de crédito, deben registrar en la subcuenta 640401 el monto de los cupos de tarjeta de crédito no utilizados.
- Además, la cuenta 640401, en el cálculo del patrimonio técnico requerido se ponderará de acuerdo al siguiente cronograma a partir de julio del año 2012:

Fecha	Factor de ponderación
julio 2012	0,10
octubre 2012	0,25
enero 2013	0,50
abril 2013	0,75
julio 2013	1,00

Tabla 1.2: Factor de ponderación para la cuenta 640401

las instituciones emisoras de TC se encuentran obligadas a desarrollar modelos matemáticos que les permitan responder las siguientes inquietudes:

1. ¿Qué características debe cumplir un individuo para ser considerado como un tarjeta-habiente Activo?.
2. ¿Cuál es el monto óptimo del cupo que se debe asignar a la tarjeta de crédito?.

Puesto que de acuerdo a la Tabla 1.2, a partir de julio del 2013 el factor de ponderación es de 1,00 (100 %), es decir si un individuo posee una TC con un cupo de 2.000 dólares, pese a que dicha tarjeta registre únicamente un monto de consumo de 200 dólares, la institución está obligada a considerar el 100 % del monto del cupo no utilizado en este caso 1.800 dólares en la suma total de activos y contingentes ponderados por riesgo y establecer un patrimonio técnico requerido del 9 %, provocando que la entidad tenga que establecer como patrimonio técnico un monto demasiado elevado de dinero, pese a que los consumos mediante la TC sean mínimos o peor aún cuando la tarjeta permanezca inactiva; el dinero que se mantendría como patrimonio técnico requerido permanecería en amortización ocasionando que la entidad pierda oportunidades en otro tipo de inversiones, a esto se suman los gastos operativos que genera mantener una tarjeta vigente a pesar de que no se la active.

En el presente trabajo abordamos la primera inquietud, es por esta razón que la finalidad del presente estudio es desarrollar una metodología analítica basada en pruebas de bondad de ajuste no paramétricas, medidas de divergencia, árboles de decisión y regresión logística para generar un modelo estadístico que permita estimar

la probabilidad que tiene un individuo de activar la tarjeta de crédito dada la apertura en una ventana de tiempo determinada.

La definición de activación puede variar de acuerdo al fenómeno de estudio, en este caso emplearemos dos definiciones y elegiremos la que mejor se pueda explicar con la información disponible.

- La primera consiste en definir a un sujeto como activa considerando el tiempo transcurrido desde la apertura hasta que realice el primer consumo.
- La segunda, en cambio, en considerar la frecuencia de consumos y el tiempo entre consumos en un periodo de tiempo fijo.

Una consideración extra que se realizará es que el monto de consumo debe ser suficiente para cubrir los costos operativos de la apertura y mantenimiento de la tarjeta y además representar cierta rentabilidad para la entidad.

Una vez desarrollada una metodología analítica surge de inmediato la cuestión de si esta metodología pueden ser implementada en algún lenguaje de programación determinado. Justamente en la segunda parte de este trabajo se pretende implementar un algoritmo en el software estadístico R, que realice automáticamente cada uno de sus pasos, con el fin de optimizar su tiempo de ejecución. El algoritmo que generaremos recibirá como parámetro de entrada nuestra base de datos con la variable dependiente y las explicativas, presentará el modelo obtenido y los resultados necesarios para validarlo.

R es un software estadístico de distribución libre que maneja programación orientada a objetos; al ser un software de distribución libre presenta las siguientes características:

- Puede ser distribuido y compartido de manera gratuita.
- Permite visualizar los algoritmos implementados y modificarlos a conveniencia.

A pesar de no tener una interfaz amigable al usuario, la facilidad de implementación y generación de algoritmos ha facilitado a que en la actualidad R sea un software muy empleado en diversas áreas tales como:

- Minería de datos.
- Clasificación de patrones.
- Estadística espacial.
- Programación lineal.
- Generación de gráficos estadísticos de alta calidad, etc.

Cabe recalcar que la metodología que implementaremos es similar a la metodología aplicada en la construcción modelos Credit Scoring, es decir que el algoritmo que generaremos tendrá la capacidad de adaptarse y utilizarse en la generación de un modelo de este tipo.

1.1. Modelos de activación.

Los modelos de activación de un determinado bien o servicio permiten pronosticar si un determinado prospecto (potencial usuario) puede convertirse en un cliente activo, es decir, que haga uso de dicho bien o servicio, en una ventana de tiempo determinada, considerando su comportamiento histórico. Este tipo de modelos son muy utilizados en las instituciones financieras y de prestación de servicios.

Dentro de la banca, entre las principales aplicaciones tenemos el predecir si un prospecto de tarjeta de crédito (TC) puede convertirse en un tarjeta-habiente activo, es decir, dado que se le concedió la tarjeta realice varios o al menos un consumo con la misma en un periodo de tiempo establecido.

Es importante predecir el comportamiento del cliente puesto que resulta menos costoso para una institución rechazar la apertura de la tarjeta que aprobarla y que el cliente no haga uso de la cuenta [Bordeleau, 2009].

Considerando el trabajo de [Parr Rud, 2001], tenemos que un método para construir un modelo de activación de tarjetas de crédito se basa en el desarrollo previo de dos modelos:

1. **Modelo de respuesta (Modelo de Originación de TC):** El modelo de respuesta aborda el problema de predecir si luego de una oferta inicial el cliente acepte la apertura de la tarjeta de crédito.

2. **Modelo de activación dada la respuesta:** En cambio el modelo de activación dada la respuesta aborda el problema de predecir si el cliente activará la tarjeta dada la apertura de la misma en una ventana de tiempo determinada.

La probabilidad de activación desde que se realiza la oferta de la tarjeta viene dada por el producto de los dos modelos anteriores. En nuestro caso nos centraremos en analizar el segundo modelo, es decir estudiaremos la activación dada la apertura o concesión de la tarjeta en un periodo de tiempo determinado.

El ajuste y la calidad de discriminación y predicción de un modelo de activación se basa principalmente en dos aspectos importantes, la técnica utilizada y la calidad de la información histórica disponible. En lo que se refiere a la información utilizada en la generación de un modelo de activación de tarjeta de crédito podemos diferenciar dos tipos de base de datos, la transaccional y la de ofertas históricas:

- **Base de datos transaccional:** Dentro de la base de datos transaccional tenemos la información sobre el historial de los consumos realizados mediante la tarjeta. Esta base de datos contiene principalmente variables tales como el número de consumos, monto de los consumos, fechas de los consumos, tipo de consumo, con esta información podemos generar un sin número de variables, las cuales generalmente son las más predictivas según afirman [Buckinx et al., 2006], entre ellas podemos enumerar las siguientes: frecuencia de consumos en los últimos 12 meses, número mínimo o máximo de consumos realizados los últimos 3 meses, porcentaje de las tarjetas que registran consumo, monto promedio de consumo los últimos 6 meses, tiempo entre consumos (tiempo inter-consumo), etc.

  Una base de datos transaccional presenta el inconveniente de ser muy difícil de generar y de emplear, debido a que en esta base se registran todas las operaciones realizadas con la tarjeta, es decir pueden existir múltiples registros

para una misma tarjeta.

- **Base de datos de ofertas históricas:** La base de datos de ofertas históricas contiene información relacionada con ofertas de tarjetas realizadas a clientes o a prospectos, por ejemplo el número de ofertas de tarjetas realizadas en los últimos 12 meses, el tiempo transcurrido desde la primera o última oferta son variables históricas que podemos encontrar en esta base de datos.

  Al igual que la base de datos transaccional, este tipo de información es bastante predictiva en la partición de los individuos en los conjuntos Activa (sujetos que activan la tarjeta) y No Activa (sujetos que no activan la tarjeta), pues un individuo al cual se le ha realizado ofertas mensuales en los últimos tres meses es menos propenso a activar la tarjeta que un individuo al que se le realiza una oferta por primera vez.

Además de la información mencionada anteriormente, es indispensable contar con variables socio demográficas tales como: edad, estado civil, género, ingresos, región y con variables asociadas al comportamiento crediticio en la institución y fuera de ella, tales como: deuda, cantidades de operaciones con vencidos en los diferentes tipos de crédito, número de tarjetas vigentes, tarjetas que registren algún consumo, etc., estas variables generalmente son proporcionadas por burós de crédito[5].

La información que describimos anteriormente es generada como se muestra en la Figura 1.3. En la ventana de comportamiento constituida por los últimos 36 meses anteriores al punto de observación, se generan las variables de las bases de datos transaccional y de ofertas históricas así como variables asociadas al comportamiento crediticio y variables socio demográficas.

Es necesario recalcar que dependiendo de las leyes de cada país, únicamente se emplea información histórica de un número de meses determinado, para el caso del Ecuador la Superintendencia de Bancos y Seguros exige únicamente considerar 36

[5]Los burós de crédito son entidades encargadas de proporcionar información sobre el historial crediticio y financiero de un determinado sujeto.

meses de historial.

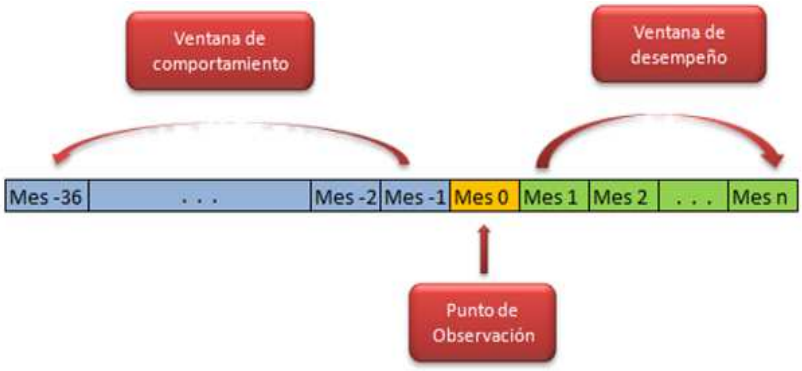

Figura 1.3: Generación información histórica

A partir de esta información se construyen los modelos de activación, los cuales se pueden diferenciar dependiendo del fenómeno que se quiera predecir, por ejemplo se pueden generar modelos que permitan:

- Pronosticar la probabilidad de realizar el primer consumo en los tres meses posteriores a la apertura (punto de observación).
- Pronosticar la probabilidad de realizar un número de consumos $c \in [n, m]$, con $n, m \in \mathbb{N}$, en periodos de tiempo regulares en los próximos 12 meses a la apertura.
- Pronosticar la probabilidad de consumir un porcentaje $p \in [0, 1]$ del cupo que registre la tarjeta, etc.

Generalmente las técnicas estadísticas más utilizadas en la generación de un modelo de activación son los árboles de decisión y la regresión logística, es por esta razón que para estimar la probabilidad de activación necesitamos un conjunto de variables independientes $X_1, X_2, ..., X_p$ que explicarán una variable dependiente binaria Y que tomará el valor de 1 si el individuo es catalogado como Activa y 0 si es catalogado como No activa, dependiendo de la definición de activación que se utilice.

En el presente estudio construiremos un modelo de activación de tarjetas de crédito para una institución perteneciente al Sistema Financiero Regulado por la Superintendencia de Bancos y Seguros (SBS), la cual tiene dentro de sus actividades

crediticias principales la emisión de tarjetas. Los datos necesarios para el desarrollo del modelo para fines académicos serán proporcionados por la institución y por burós de crédito, datos adicionales serán descargados vía web de la página oficial de la SBS.

Contaremos con información histórica de las entidades del Sistema Financiero Regulado por la Superintendencia de Bancos y Seguros (SBS), de las entidades reguladas por la Superintendencia de Economía Popular y Solidaria (SEPS), y de casas comerciales.

La información disponible para la construcción del modelo será muy completa, pues además de la información en la institución (comportamiento en la institución), se contará con información del Sistema Crediticio Ecuatoriano (comportamiento fuera de la institución), esto nos permitirá generar un modelo altamente predictivo, lo cual no sería posible considerando únicamente la información dentro de la institución.

Cabe recalcar que para el presente trabajo no fue posible contar con información de ofertas de tarjetas históricas, ni tampoco con información de transacciones de TC detallada por operación, debido a que representa un elevado costo para la institución generarla o simplemente la entidad no disponía de la misma. La mayoría de información disponible se encuentra consolidada ya sea por tarjeta, por individuo, o mes, esto con el propósito por tener una mayor facilidad al momento de que dicha información sea generada.

La muestra con la cual desarrollaremos el estudio serán todas las tarjetas que registran como fecha de apertura los meses enero 2012, abril 2012, julio 2012 y octubre 2012, seleccionaremos la muestra de esta manera con el fin de abarcar el comportamiento anual y eliminar posibles estacionalidades como por ejemplo el inicio de clases, pago de utilidades, festividades, etc.

La estructura del presente estudio será la siguiente, en el capítulo 2, describiremos los aspectos teóricos necesarios para comprender la metodología utilizada en la

generación del modelo de activación de TC. En el capítulo 3, describiremos la metodología y desarrollaremos el modelo de activación con la información disponible. En el capítulo 4, implementaremos y explicaremos un algoritmo en R que realice automáticamente cada uno de los pasos de la metodología utilizada en la generación del modelo; para finalmente, en el capítulo 5, realizar la comparación entre los resultados obtenidos en el capítulo 3 y 4. En el capítulo 6 presentaremos los hallazgos más importantes realizados durante el desarrollo del trabajo.

Capítulo 2

Marco teórico.

En este capítulo estudiaremos las nociones y conceptos teóricos necesarios para comprender la metodología utilizada en la construcción del modelo de activación de tarjetas de crédito, empezaremos describiendo ciertos estadísticos e índices que miden la divergencia entre las distribuciones de probabilidad de dos variables aleatorias, la técnica de los árboles de decisión, para finalmente estudiar el modelo de regresión logística, en el cual se fundamenta el modelo de activación.

2.1. Medidas de separación o divergencia.

Es importante en el desarrollo de un modelo estadístico, conocer el tipo y la calidad de información disponible, debido a que la calidad del modelo dependerá, en gran medida, de la información utilizada en la construcción del mismo. En el caso de los modelos de activación de tarjetas de crédito se requiere predecir la probabilidad de que un individuo active la tarjeta, es por esta razón que se necesita que la información utilizada permita identificar, de la mejor manera posible, las características de quienes mantienen activa la tarjeta y de quienes no la mantienen.

Para la construcción de este tipo de modelos, en ocasiones se cuenta con una gran cantidad de información (variables explicativas), esto genera problemas al momento de analizar y procesar la misma, de ahí surge la necesidad de realizar un filtrado previo de las variables explicativas, puesto que podría ser suficiente considerar un menor número de variables y obtener resultados similares a los que se conseguirían

si se analizara toda la gama de variables explicativas disponibles.

En esta sección nos centraremos en estudiar algunos índices y estadísticos, mismos que indican qué tanto se diferencian (divergen) las distribuciones de individuos Activa y No Activa para cada variable explicativa, conocidos como medidas de separación o divergencia, es decir a través de estas medidas podemos conocer el poder predictivo de cada variable, de esta manera podemos seleccionar aquellas con el mayor poder predictivo y realizar un análisis más exhaustivo únicamente sobre ellas.

Las medidas de separación pueden ser empleadas en diversas etapas del desarrollo de los modelos de clasificación binaria, tales como:

- **Selección de variables:** Cuando se requiere obtener un subconjunto de variables explicativas con un mayor poder de predicción.
- **Segmentación:** Cuando existe la necesidad de segmentar la población en varios grupos debido al diferente comportamiento dentro de cada uno; por ejemplo, podría suceder que los habitantes de la Costa registren una tasa de activación más elevada que la de los habitantes de cualquier otra región, de ser este el caso se debe desarrollar un modelo para cada uno de los grupos.
- **Validación:** Cuando verificamos la calidad de discriminación y predicción del modelo generado.
- **Monitoreo:** Cuando analizamos el funcionamiento del modelo en el tiempo.

2.1.1. Prueba de Kolmogorov-Smirnov para dos muestras (KS).

Utilizando el trabajo de [Arnold and Emerson, 2011], procedemos a describir la prueba de Kolmogorov-Smirnov (1933) para dos muestras aleatorias. El test de Kolmogorov-Smirnov es una prueba de bondad de ajuste, mediante la cual se contrasta la hipótesis de si dos muestras aleatorias independientes provienen de distribuciones continuas idénticas; es una prueba del tipo no paramétrico debido a que no es necesario realizar suposiciones a priori sobre la distribución de los datos.

Para calcular el estadístico KS de la prueba, se utiliza la función de distribución empírica acumulada, es por esta razón, que para describir la prueba KS empezaremos definiendo la distribución de acumulación empírica de una variable aleatoria.

Definición 1. ***Distribución de acumulación empírica (ecdf).*** *Consideremos una muestra de tamaño n, $x_1, x_2, ..., x_n$ de una variable aleatoria X, definimos la distribución acumulada empírica de X mediante la expresión:*

$$ecdf(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1 & Si\ x_i \leq x. \\ 0 & caso\ contrario. \end{cases} \tag{2.1}$$

ahora describimos la prueba de Kolmogorov-Smirnov para dos muestras aleatorias, para ello consideremos: $x_1, x_2, ..., x_{n_1}$ una muestra de tamaño n_1 de una variable aleatoria continua X con función de distribución acumulada F_1; $y_1, y_2, ..., y_{n_2}$ una muestra de tamaño n_2 de una variable aleatoria continua Y con función de distribución acumulada F_2. En la prueba KS se contrastan las hipótesis:

$$\begin{cases} H_0 : F_1(x) = F_2(x)\ \forall x \\ H_1 : F_1(x) \neq F_2(x) \end{cases} \tag{2.2}$$

el estadístico KS empleado para contrastar la hipótesis nula H_0 está basado en la utilización de la función de distribución acumulada empírica de X y de Y, y su valor se obtiene mediante la expresión:

$$KS = max_x|ecdf_1(x) - ecdf_2(x)| \tag{2.3}$$

donde $ecdf_1$ denota la función de acumulación empírica de X, y $ecdf_2$ la función de distribución empírica de Y. La hipótesis nula H_0 se rechaza si el estadístico KS es mayor que el valor crítico KS_α, para un nivel de significancia α dado. [Massey, 1951] presenta una tabla de valores críticos para diferentes tamaños de muestra.

De la expresión (2.3), podemos decir que el estadístico KS es la distancia máxima entre $ecdf_1$ y $ecdf_2$ y su valor oscila entre 0 y 1, donde valores cercanos a 0 indican que las distribuciones de X y de Y son idénticas, y valores cercanos a 1 indican que las distribuciones de X y de Y difieren, es por este motivo, que el estadístico KS es utilizado como una medida de divergencia entre las distribuciones de dos variables

aleatorias continuas.

En el presente estudio, la prueba KS será utilizada para seleccionar aquellas variables que generen la mayor divergencia entre las distribuciones de los individuos Activa y No activa, es decir seleccionaremos las variables que presenten un estadístico KS superior a un valor especifico mayor que cero.

Para una variable continua X arbitraria, calcularemos el estadístico KS como se describe a continuación. Seleccionaremos los valores correspondientes a los individuos etiquetados como Activa y formaremos una nueva variable X_A y con los valores de la variable X correspondientes a los individuos No Activa formaremos la variable X_{NA}, nuestro objetivo es comparar las distribuciones de X_A y X_{NA} y seleccionar las variables que presenten un mayor valor del estadístico KS, para de esta manera obtener un subconjunto de las variables explicativas con el mayor poder predictivo, las cuales permitirán obtener una mejor partición del conjunto de individuos en Activa y No Activa.

A continuación presentamos un ejemplo de la interpretación gráfica y la forma de cálculo en R del estadístico de Kolmogorov-Smirnov, consideremos dos variables aleatorias simuladas, una con distribución exponencial y la otra con distribución normal.

$$\begin{cases} X_{NA} \to exp(1) \\ X_A \to N(3,1) \end{cases}$$

Por la forma de simulación de las variables es de esperar que el estadístico KS sea elevado. Empecemos graficando la función de densidad tanto para la variable X_A como para X_{NA}. En la Figura 2.1, observamos que existe una alta separación entre las distribuciones de las variables X_A y X_NA de sujetos Activa y No Activa respectivamente, por lo cual es de esperar obtener un estadístico KS cercano a 1.

El gráfico de la Figura 2.2, es conocido como curva de KS, la curva de KS es una instrumento de visualización empleado para ilustrar la discriminación existente entre

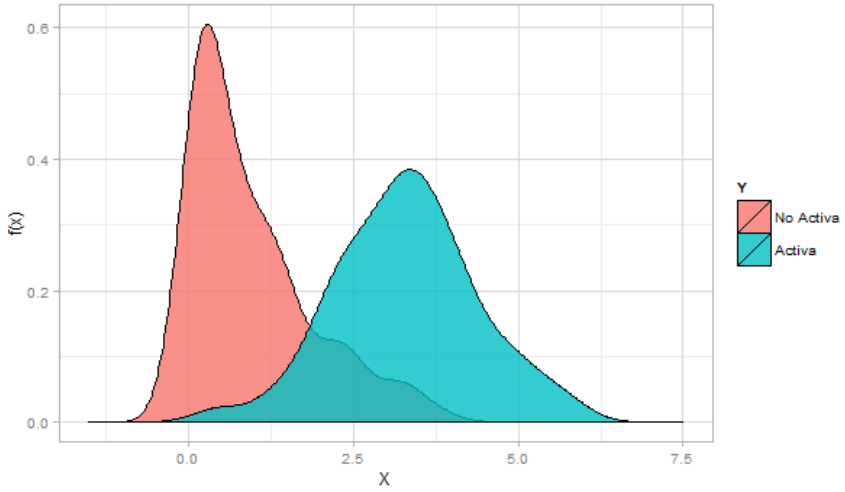

Figura 2.1: Funciones de densidad de sujetos Activa y No Activa.

las distribuciones analizadas. Calculando el estadístico KS como la diferencia máxima absoluta entre las curvas Activa y No Activa tenemos:

$$KS = |0,86 - 0,10| = 0,76$$

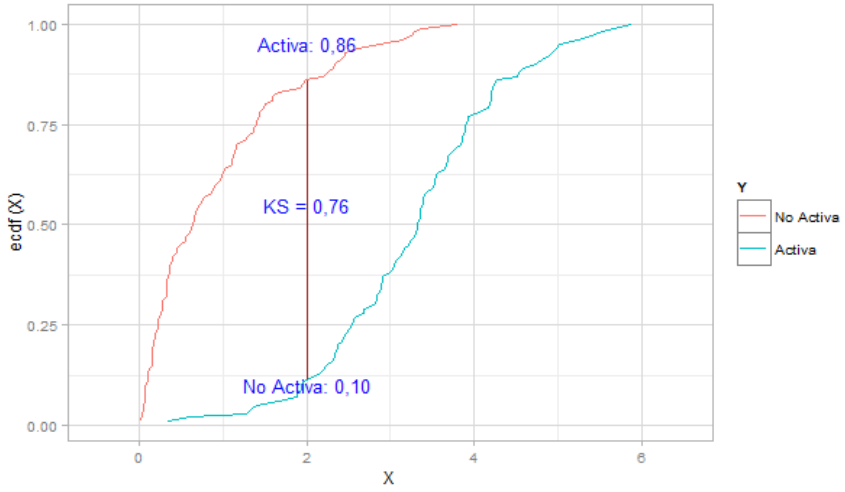

Figura 2.2: Curva de KS.

Utilizando el trabajo de [Lewis, 2013], un posible código en R para la tarea anteriormente descrita es el siguiente:

```
# Simulación de variables
 x1 <- rnorm(100,3,1)
 x2 <- rexp(100,1)
```

```
 # Cálculo KS
  ks.test(x1,x2)


  Two-sample Kolmogorov-Smirnov test
data:  x1 and x2
D = 0,76, p-value < 2,2e-16
alternative hypothesis: two-sided
```

2.1.2. Prueba de Anderson Darling para dos muestras (AD).

Para contrastar la hipótesis (2.2), a parte del estadístico de Kolmogorov-Smirnov existe una prueba introducida por Anderson-Darling (1952, 1954).

La prueba de Anderson-Darling (AD) es en general más potente que la prueba de Kolmogorov-Smirnov, una de las razones es que el estadístico KS se calcula como la distancia máxima entre las funciones de acumulación empíricas, es decir no se considera el comportamiento de las colas de las distribuciones.

[Engmann and Cousineau, 2011] realizan un análisis muy completo acerca de la diferencia entre el uso de la prueba KS y la prueba AD, después de una serie de comparaciones los autores concluyen que la prueba AD es en general más potente que la prueba KS.

La prueba de Anderson-Darling se puede extender a la verificación de si k muestras aleatorias independientes provienen de distribuciones idénticas, en nuestro caso es suficiente con describir la prueba AD para el caso de dos muestras independientes.

[Scholz and Stephens, 1987] describen a detalle el cálculo del estadístico AD utilizado para contrastar la hipótesis (2.2), al igual que el estadístico KS su cálculo se basa en la diferencia entre los valores de las funciones de distribución acumuladas empíricas. Para describir su ecuación al igual que en la descripción de la prueba KS, consideremos $x_1, x_2, ..., x_{n_1}$ una muestra de tamaño n_1 de una variable aleatoria continua X con función de distribución acumulada F_1; $y_1, y_2, ..., y_{n_2}$ una muestra de

tamaño n_2 de una variable aleatoria continua Y con función de distribución acumulada F_2, teóricamente el estadístico AD está dado por la expresión (2.4).

$$AD = \frac{n_1.n_2}{N} \int_{-\infty}^{\infty} \frac{[F_1(x) - F_2(x)]^2}{H_N(x)(1 - H_N(x))} dH_N(x) \tag{2.4}$$

donde:

$$N = n_1 + n_2 \tag{2.5}$$

$$H_N(x) = \frac{n_1 F_1(x) - n_2 F_2(x)}{N} \tag{2.6}$$

En el trabajo desarrollado por [Scholz and Stephens, 1987], se propone utilizar la fórmula computacional (2.7), como una aproximación de (2.4).

$$AD = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{[NM_{ij} - jn_i]^2}{j(N-j)} \tag{2.7}$$

con $k = 2$. Ordenando conjuntamente los valores de las variables X y Y tenemos la muestra ordenada $Z_1 < Z_2 < ... < Z_N$, y M_{ij} es el número de observaciones de la muestra i que son menores o iguales a Z_j.

Considerando el trabajo desarrollado por [Lewis, 2013], a continuación presentamos un ejemplo del cálculo del estadístico AD en R.

```
# Simulación de variables
 x1 <- rnorm(100,3,1)
 x2 <- rexp(100,1)
# Cálculo KS
 ad.test(x1,x2)


  Anderson-Darling k-sample test.
Number of samples:  2
Sample sizes:  100, 100
Number of ties: 0
Mean of  Anderson-Darling  Criterion: 1
Standard deviation of  Anderson-Darling  Criterion: 0,75419
T.AD = ( Anderson-Darling  Criterion - mean)/sigma
```

```
Null Hypothesis: All samples come from a common population.
    AD    T.AD  asympt. P-value
   50,074  5,068  5,5602e-28
```

2.2. Medidas de asociación.

Las medidas de asociación son básicamente utilizadas para medir el poder predictivo de las variables categóricas consideradas como candidatas a formar parte del modelo, a través de estas medidas es posible realizar un filtrado previo de las variables categóricas y así seleccionar aquellas que permitan generar un modelo de predicción con un mejor ajuste y con un elevado poder predictivo.

El cálculo de ciertas medidas se basa en analizar la diferencia de los porcentajes de individuos Activa y No Activa en las distintas categorías de la variable respecto a los porcentajes en la población total. Otras medidas en cambio analizan la diferencia de los porcentajes de individuos Activa y No Activa en una categoría respecto a los porcentajes en las categorías restantes. A continuación describimos algunas de las medidas más utilizadas en la práctica.

2.2.1. Prueba de independencia Ji-Cuadrado (χ^2).

En este caso utilizaremos la prueba de independencia Ji-Cuadrado χ^2 para probar si dos variables aleatorias categóricas son independientes. El cálculo del estadístico χ^2 de la prueba está basado en la diferencia entre las frecuencias observadas y las frecuencias esperadas dada la independencia.

Nuestro propósito, en el presente estudio, es utilizar el estadístico χ^2 para estudiar la independencia entre la variable dependiente binaria Y : Activa/No Activa y una variable cualitativa arbitraria X con $C_1, C_2, ...C_p$ categorías. Por lo tanto para describir la prueba χ^2, consideremos las variables cualitativas: X con p categorías y Y con 2 categorías, en esta prueba se contrastan las hipótesis:

$$\begin{cases} H_0 : X, Y \, son \, independientes \\ H_1 : X, Y \, no \, son \, independientes \end{cases} \tag{2.8}$$

utilizando el esquema de contingencia de la Tabla 2.1, para estudiar la relación entre X y Y, tenemos que el valor del estadístico χ^2 asociado a la prueba puede ser calculado mediante la ecuación (2.9).

$$\chi^2 = \sum_{i=1}^{p} \frac{(a_i - \widehat{a}_i)^2}{\widehat{a}_i} + \frac{(na_i - \widehat{na}_i)^2}{\widehat{na}_i} \tag{2.9}$$

donde, a_i, na_i son las frecuencias observadas de sujetos Activa y No Activa respectivamente, $n_i = a_i + na_i$, y $\widehat{a}_i$ y $\widehat{na}_i$ son las frecuencias esperadas de sujetos Activa y No Activa, respectivamente, en la categoría C_i, definidas mediante las expresiones (2.10) y (2.11).

$$\widehat{a}_i = \frac{(a_i + na_i)A}{n} \tag{2.10}$$

$$\widehat{na}_i = \frac{(a_i + na_i)NA}{n} \tag{2.11}$$

donde, A es el número total de sujetos Activa, NA es el número total de sujetos No Activa, y $n = A + NA$.

Y \ X	C_1	.	.	.	C_i	.	.	.	C_p	**Total**
Activa	a_1	.	.	.	a_i	.	.	.	a_p	A
No Activa	na_1	.	.	.	na_i	.	.	.	na_p	NA
Total	n_1	.	.	.	n_i	.	.	.	n_p	n

Tabla 2.1: Esquema de contingencia variable dependiente y variable explicativa

En la mayoría de softwares estadísticos, el estadístico χ^2 es utilizado como una medida de la relación entre la variable dependiente Y y la variable independiente X. Se debe recalcar que considerando el esquema de contingencia de 2 filas y p columnas de la Tabla 2.1, se tiene que el estadístico χ^2 verifica la desigualdad siguiente:

$$0 \leq \chi^2 \leq n \, min\{2 - 1, p - 1\}$$

$$0 \leq \chi^2 \leq n \, min\{1, p - 1\}$$

como $p \geq 2$, se tiene que,

$$0 \leq \chi^2 \leq n\,(p-1) \tag{2.12}$$

es por esta razón que en nuestro caso, como una medida de relación entre variables cualitativas utilizaremos el siguiente coeficiente conocido como ***coeficiente de contingencia de Pearson***:

$$CCP = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{2.13}$$

el mismo que verifica que la desigualdad $0 \leq CCP < 1$, donde valores cercanos a 0 indicarán independencia entre las variables y valores cercanos a 1 indicarán que existe relación entre las variables analizadas.

Probemos ahora que en realidad se verifica la desigualdad $0 \leq CCP < 1$. Puesto que $0 \leq \chi^2 \leq n\,(p-1)$, tenemos,

$$0 \leq \chi^2 \leq np - n$$

$$n \leq \chi^2 + n \leq np$$

$$\frac{1}{n} \geq \frac{1}{\chi^2 + n} \geq \frac{1}{np}$$

$$-1 \leq -\frac{n}{\chi^2 + n} \leq -\frac{1}{p}$$

$$0 \leq 1 - \frac{n}{\chi^2 + n} \leq 1 - \frac{1}{p}$$

$$0 \leq \frac{\chi^2}{\chi^2 + n} \leq \frac{p-1}{p}$$

$$0 \leq \sqrt{\frac{\chi^2}{\chi^2 + n}} \leq \sqrt{\frac{p-1}{p}}$$

$$0 \leq CCP \leq \sqrt{\frac{p-1}{p}}$$

como $\frac{p-1}{p} < 1$, se tiene que $0 \leq CCP < 1$.

Como podemos notar en la demostración anterior la cota superior de CPP, depende del número de categorías de la variable explicativa X, para evitar esto algunos autores

prefieren utilizar el siguiente coeficiente de contingencia corregido para medir la dependencia:

$$CCP_{corr} = \sqrt{\frac{p}{p-1}}\sqrt{\frac{\chi^2}{\chi^2+n}}$$

el cual verifica que $0 \leq CCP_{corr} \leq 1$

2.2.2. Valor de Información (VI).

En problemas de clasificación binaria (Activa / No Activa), el valor de información de una variable categórica, según [Finlay, 2010], se puede considerar como una medida del poder predictivo de la variable; específicamente el VI mide la diferencia entre la probabilidad de que la variable tome el valor de la categoría i, dado que el individuo es Activa $(p(X = C_i|A))$ y la probabilidad de que la variable tome el valor de la categoría i, dado que el individuo es No Activa $(p(X = C_i|NA))$.

Considerando la Tabla 2.1, podemos definir el valor de información como sigue:

$$VI = \sum_{i=1}^{p} \left(\frac{na_i}{NA} - \frac{a_i}{A}\right) \ln\left(\frac{na_i/NA}{a_i/A}\right) \tag{2.14}$$

mientras mayor sea VI, mayor será la diferencia entre $p(X = C_i|A)$ y $p(X = C_i|NA)$. En el presente estudio se seleccionarán las variables categóricas que presenten el mayor valor de información.

Es importante señalar que un inconveniente que presenta el VI es que no considera el porcentaje de sujetos en cada categoría respecto a la población total, es por este motivo que se propone una nueva medida de información, la cual en la práctica ha resultado ser de mucha importancia, esta medida se describe a continuación.

2.2.3. Diferencia de Información por categoría (DIC).

En el presente trabajo estudiamos un problema del tipo Bernoulli, en el cual, queremos pronosticar la probabilidad de *Éxito: Individuo Activa*, es por esta razón, que para una variable categórica nos centramos únicamente en estudiar la diferencia entre el

porcentaje de éxito en cada categoría y el porcentaje de éxito en la población. El índice (DIC) se puede considerar como una medida del poder predictivo de la variable pues su cálculo se basa en la suma de las diferencias ponderadas entre la probabilidad de que un individuo sea Activa y las probabilidades de que sea Activa dado que la variable toma el valor de los atributos de la variable categórica.

Procedamos ahora a explicar el cálculo del índice (DIC) para ello consideremos la Tabla 2.1, y p_A como el porcentaje de individuos Activa en la población, el índice (DIC) se define mediante la igualdad:

$$DIC = \sqrt{\sum_{i=1}^{p} \frac{1}{n_i}\left(\frac{a_i}{n_i} - p_A\right)^2} \tag{2.15}$$

2.3. Árboles de decisión.

En minería de datos, una de las técnicas más utilizadas en problemas de clasificación y predicción es el árbol de decisión, conocido también como algoritmo de particionamiento recursivo (RPA). El árbol de decisión es una técnica no paramétrica puesto que no hay la necesidad de realizar suposiciones a priori sobre la distribución del conjunto de datos analizado.

Para describir esta técnica de clasificación consideremos una variable dependiente binaria $(Activa/No\ Activa)$ y una variable explicativa (cualitativa o cuantitativa), la técnica del árbol de decisión es un algoritmo recursivo que consiste en particionar la población de estudio en segmentos homogéneos (un segmento se considera homogéneo si los individuos pertenecientes al segmento tienen aproximadamente la misma probabilidad de ser catalogados como Activa o No Activa) mediante la utilización de reglas de partición basadas en los valores que tome la variable explicativa.

El proceso iterativo que se sigue para generar los segmentos es el siguiente: Inicialmente se particiona la población de dos subconjuntos homogéneos, luego cada uno de estos subconjuntos es particionado nuevamente en dos subconjuntos más

homogéneos, el proceso es repetido recursivamente y termina si el subconjunto presenta una cantidad de individuos menor o igual a la mínima requerida (criterio de parada), finalmente se establece el tipo del subconjunto $(Activa/No\ Activa)$ dependiendo de su distribución de individuos Activa y No Activa respecto a la distribución en el conjunto inicial (criterio de asignación). A continuación se explican los criterios utilizados en la ejecución del algoritmo:

1. **Criterio de partición.** Para establecer los valores de corte de la variable explicativa que definirán los segmentos se utiliza el método de particionamiento CHAID, el cual se basa en el estadístico χ^2 para evaluar la dependencia entre la variable dependiente y la variable categórica construida en base a los criterios de partición generados.

2. **Criterio de parada.** Un subconjunto se particiona solo si su porcentaje de individuos es mayor a un porcentaje previamente establecido, por ejemplo $3\,\%$ del total de la población.

3. **Criterio de asignación.** Luego de verificarse el criterio de parada, los subconjuntos finales obtenidos se conocen como terminales, los subconjuntos terminales en nuestro caso serán de dos tipos:

    - **Activa.** Si el porcentaje de individuos Activa en el subconjunto terminal es mayor que el porcentaje de individuos Activa en el conjunto inicial.
    - **No Activa.** Si el porcentaje de individuos No Activa en el subconjunto terminal es mayor que el porcentaje de individuos No Activa en el conjunto inicial.

En el presente estudio emplearemos los árboles de decisión para identificar y construir características (variables dummys y probabilidades de activa) que permitan generar una partición de la población en segmentos homogéneos. A continuación presentamos un ejemplo de un árbol de decisión, en el cual explicaremos cómo funciona el método de partición CHAID, y cómo utilizaremos esta técnica para generar variables binarias conocidas como dummys, y variables basadas en el porcentaje de individuos Activa en cada subconjunto terminal (probabilidades de activa), mediante

las cuales estableceremos una mejor partición de la población en individuos Activa y No Activa.

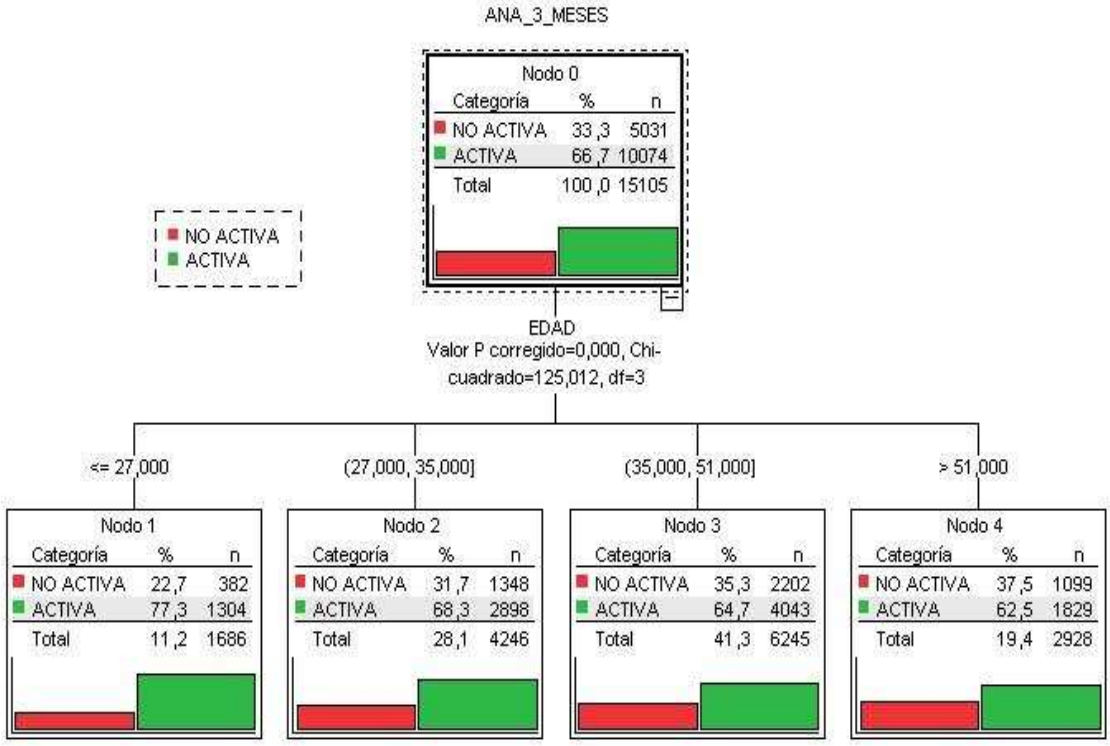

Figura 2.3: Árbol de decisión para la variable Edad.

En la Figura 2.3, presentamos un ejemplo de un árbol de decisión construido para la variable *X:Edad*, en este gráfico observamos que cada subconjunto generado es representado mediante un nodo, el nodo inicial es conocido como nodo padre, el cual está constituido por la población total, y cada nodo generado es conocido como nodo hijo, una vez verificado el criterio de parada los nodos obtenidos se conocen como nodos terminales.

El Nodo 1 formado por los sujetos con una edad menor o igual a 27 años presenta un porcentaje de individuos Activa de $77,30\,\%$; el Nodo 2 formado por los sujetos con una edad mayor a 27 y menor o igual que 35 años presenta un porcentaje de individuos Activa de $68,30\,\%$; el Nodo 3 formado por los sujetos con una edad mayor a 35 y menor o igual que 51 años presenta un porcentaje de individuos Activa de $64,70\,\%$; finalmente tenemos el Nodo 4 formado por los sujetos con una edad mayor a 51 años presenta un porcentaje de individuos Activa de $62,50\,\%$.

Ahora explicaremos el proceso de construcción de las variables dummys, el objetivo de la construcción de las variables dummys es el generar características con un mayor poder predictivo en la clasificación de los individuos en Activa y No Activa. La generación de la variable dummy se basa en analizar la diferencia entre el porcentaje de individuos Activa y No Activa en los nodos hijos respecto a los porcentajes en el nodo padre. En este caso, para el Nodo 1 tenemos un $77,30\,\%$ de sujetos Activa este porcentaje es mayor que el de sujetos Activa en el nodo padre $66,70\,\%$, es por esta razón que el nodo 1 es etiquetado como un nodo de tipo Activa.

Matemáticamente, las variables probabilidad de activa ($prba\_Edad$) y dummy (d_X) están dadas por las siguientes expresiones:

$$prba\_Edad = \begin{cases} 0,773 & \text{Si } Edad \leq 27. \\ 0,683 & \text{Si } 27 < Edad \leq 35. \\ 0,647 & \text{Si } 35 < Edad \leq 51. \\ 0,625 & \text{Si } Edad > 51. \end{cases}$$

$$d\_Edad = \begin{cases} 1 & \text{Si } Edad \leq 35. \\ 0 & \text{Si } Edad > 35. \end{cases}$$

puesto que nuestro objetivo es pronosticar la probabilidad de que un sujeto sea Activa, nos centraremos en construir variables dummys del tipo Activa.

2.4. Regresión logística.

En esta sección resumiremos la descripción de la regresión logística presentada en [Castro, 2008]. La regresión logística es una de las técnicas paramétricas más utilizadas para predecir una variable categórica mediante un conjunto de variables explicativas, en este caso estudiaremos la regresión logística múltiple para predecir los valores de una variable dependiente binaria (Activa/No Activa) que toma el valor de 1 para las observaciones etiquetadas como Activa y 0 para las observaciones etiquetadas como No Activa.

Para describir esta técnica consideremos una muestra de n individuos, $X_2, X_3, ..., X_p$, $p-1$ variables explicativas o regresores, y una variable dependiente Y. Considerando como y_i el valor de la variable Y en el individuo i, definimos y_i como sigue:

$$y_i = \begin{cases} 1 & \text{Si el individuo i es etiquetado como Activa.} \\ 0 & \text{Si el individuo i es etiquetado como No Activa.} \end{cases} \tag{2.16}$$

notando como $\pi_i = Pr(y_i = 1)$ y $1 - \pi_i = Pr(y_i = 0)$, nuestro objetivo es establecer una relación entre π_i y los $p-1$ regresores, es decir:

$$\pi_i = f(x_{i2}, x_{i3}, ..., x_{ip}; \beta_1, \beta_2, ..., \beta_p)$$

donde f es una función y x_{ij} es el valor de la variable j en el individuo i, $\beta_1, \beta_2, ..., \beta_p)$ son constantes desconocidas que deben ser estimadas.

Si f es lineal respecto a las variables explicativas tenemos el modelo siguiente:

$$\pi_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} + u_i \tag{2.17}$$

El modelo (2.17) es conocido como modelo lineal, este modelo no es adecuado para predecir π_i pues el valor de su estimación $\hat{\pi}_i$ puede resultar fuera del intervalo $[0, 1]$. Para asegurar que $\hat{\pi}_i \in [0, 1]$, se consideran funciones f las cuales tienen como rango el intervalo $[0, 1]$, una de las funciones más empleadas es la **función logística** definida como sigue:

$$\pi_i = \frac{1}{1 + exp(-\eta_i)} \quad i = 1, 2, ..., n \tag{2.18}$$

donde:

$$\eta_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip}$$

Gráficamente la función logística presenta la forma de la Figura 2.4.

De la ecuación (2.18) despejando η_i tenemos:

$$\eta_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right)$$

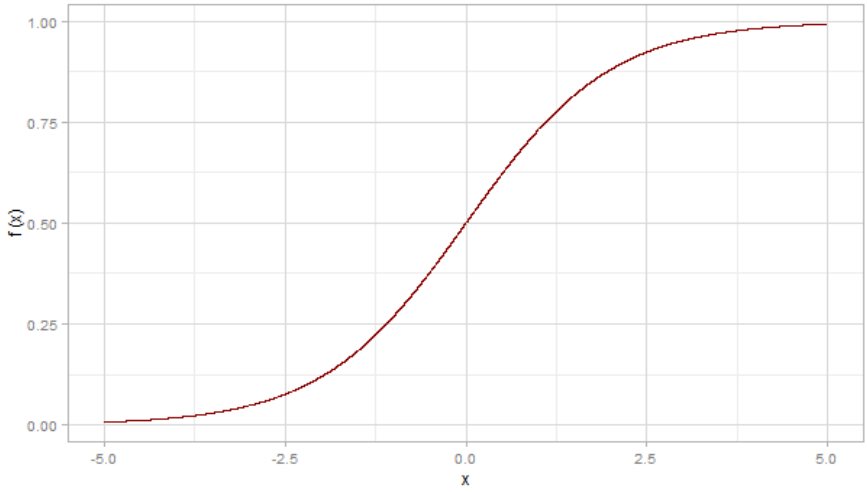

Figura 2.4: Función logística

el cual es conocido como modelo $logit$ y se nota por:

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} \tag{2.19}$$

La razón $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$, se denomina razón de probabilidades, recordando que $\pi_i = Pr(y_i = 1)$ denota la probabilidad de activación y $1 - \pi_i = Pr(y_i = 0)$ la probabilidad de no activación.

2.4.1. Validación del modelo de regresión logística.

A parte de las pruebas estadísticas clásicas utilizadas en la validación de un modelo de regresión logística, tales como pruebas sobre los coeficientes estimados, significancia del modelo, análisis de los residuos, coeficiente de determinación, etc., se utilizan varios estadísticos e indicadores adicionales.

En esta sección precisamente nos centraremos en describir los indicadores más empleados en la validación de la predicción y la discriminación de un modelo logístico.

- **Estadístico** KS**:** Este estadístico es utilizado para medir la divergencia entre las distribuciones de la probabilidad de activación estimada mediante el modelo logístico para los individuos Activa y No Activa. Se explicó a detalle en la sección (2.1.1) del capítulo 2.

- **Área bajo la curva ROC** $(AUROC)$**:** Una vez estimada la probabilidad de activación P_{Ai} para el individuo i, se procede a clasificarlo en Activa o No Activa, es decir se construye la variable dependiente $\hat{Y}$ pronosticada por el modelo, cuyo valor en el individuo i se calcula utilizando la siguiente expresión:

$$\hat{y}_i = \begin{cases} 1 & \text{Si } P_{Ai} > p_0. \\ 0 & \text{Si } P_{Ai} \leq p_0. \end{cases}$$

donde p_0 es conocido como punto de corte y es el valor mínimo de la probabilidad de activación a partir del cual un individuo es clasificado como Activa.

Utilizando el trabajo de [Siddiqi, 2006], tenemos que la curva ROC (Receiver Operating Characteristics) es una herramienta gráfica en la cual podemos visualizar el rendimiento de un modelo de clasificación, pues las coordenadas de su gráfico son las parejas ordenadas:

1. **Sensibilidad.** Razón de verdaderos positivos.
2. **Complemento de la especificidad.** Razón de falsos positivos.

que resultan de variar continuamente el punto de corte o umbral de clasificación en el intervalo [0,1].

Si un individuo positivo (Activa) es clasificado como positivo se denomina verdadero positivo, por el contrario si es clasificado como negativo (No Activa) se denomina falso positivo, en cambio si un individuo negativo es clasificado como negativo se denomina verdadero negativo, por el contrario si es clasificado como positivo se denomina falso negativo, tal como muestra el esquema de contingencia entre la variable dependiente real y la pronosticada de la Tabla 2.2, conocida como **matriz de confusión**.

Formalmente podemos definir la curva ROC como el conjunto de pares ordenados definidos por las expresiones siguientes:

Real \ Pronóstico	**Activa**	**No Activa**	**Total**
Activa	Verdaderos Positivos (vp)	Falsos Negativos (fn)	Total Positivos (tp)
No Activa	Falsos Positivos (fn)	Verdaderos Negativos (vn)	Total Negativos (tn)

Tabla 2.2: Esquema de contingencia variable dependiente real y pronosticada.

$$Sensibilidad = P[P_A > p_0 \,|\, Y = 1] \tag{2.20}$$

$$1 - Especificidad = P[P_A > p_0 \,|\, Y = 0] \tag{2.21}$$

$$ROC = \{ \, (P[\, P_A > p_0 | Y = 0\,],\, P[\, P_A > p_0 | \, Y = 1\,] \,) : p_0 \in [0,1] \} \tag{2.22}$$

Con p_0 variando en el intervalo [0,1], el punto de corte óptimo es el valor de p_0 para el cual el punto $(P[\, P_A > p_0 | Y = 0\,],\, P[\, P_A > p_0 | \, Y = 1\,] \,)$ es más próximo al punto (0,1), además este punto de corte coincide con el valor de la probabilidad de activación pronosticada en el que se maximiza el KS, verifiquemos lo anterior y procedamos a demostrar la expresión:

$$KS = max_{p_0} \, |F_{NA}(p_0) - F_A(p_0)| = max_{p_0} |sensibilidad - (1 - especificidad)|$$

$$|sensibilidad - (1 - especificidad)| = |P[P_A > p_0 | Y = 1] - P[P_A > p_0 | Y = 0]|$$

$$= |1 - P[P_A \leq p_0 | Y = 1] - (1 - P[P_A \leq p_0 | Y = 0])|$$

$$= |1 - P[P_A \leq p_0 | Y = 1] - 1 + P[P_A \leq p_0 | Y = 0]|$$

$$= |P[P_A \leq p_0 | Y = 0] - P[P_A \leq p_0 | Y = 1]|$$

$$|sensibilidad - (1 - especificidad)| = |F_{NA}(p_0) - F_A(p_0)|$$

$$max_{p_0} \, |sensibilidad - (1 - especificidad)| = max_{p_0} \, |F_{NA}(p_0) - F_A(p_0)|$$

$$max_{p_0} \, |sensibilidad - (1 - especificidad)| = max_{p_0} \, |F_{NA}(p_0) - F_A(p_0)| = KS$$

En la Figura 2.5, presentamos un ejemplo del gráfico de la curva ROC. El punto (0, 1) representa una clasificación perfecta, mientras que un punto a lo largo de

la recta $y = x$ representa una clasificación totalmente aleatoria.

Mediante la curva ROC se generan índices que miden el rendimiento de un clasificador a lo largo del rango de la probabilidad pronosticada, definiendo como rendimiento la capacidad de clasificar correctamente las observaciones. Uno de ellos es el área bajo la curva ROC conocido como $AUROC$ (área under ROC curve) el mismo que según [Fawcett, 2005] tiene la propiedad de ser equivalente a la probabilidad que tiene un clasificador de obtener una probabilidad de activación más alta para un individuo positivo que para un negativo elegidos aleatoriamente.

El valor del índice $AUROC$ al ser una porción del área del cuadrado unitario oscila entre 0 y 1, donde valores cercanos a 1 indican un alto rendimiento del modelo de clasificación. Según [Anderson, 2007], un valor para $AUROC$ de 0,0 implica que las predicciones del modelo son perfectamente erróneas, un valor de 0,5 indica que el modelo realiza un predicción aleatoria y un valor de 1 implica que el modelo realiza una predicción perfecta; y que generalmente un valor superior a 0,7 es considerado adecuado.

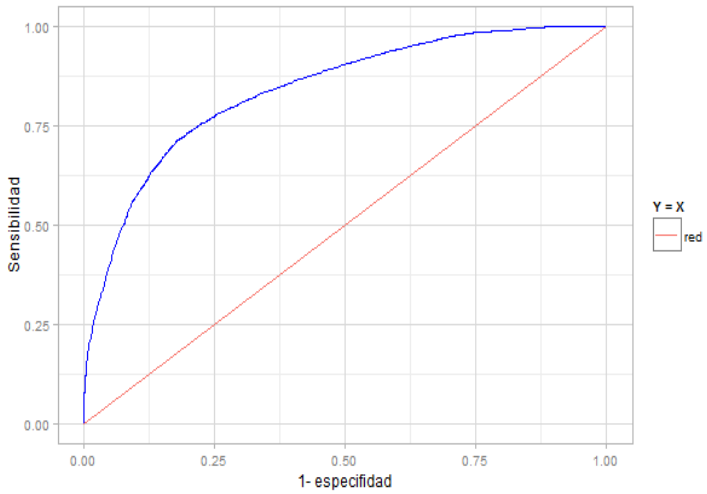

Figura 2.5: Curva ROC

- **Coeficiente de** $GINI$**:** El coeficiente de $GINI$ al igual que el indicador $AUROC$,

es una medida de qué tan bien el modelo de regresión clasifica a individuos positivos y negativos cuando el punto de corte varía a lo largo del rango de la probabilidad pronosticada, oscila entre 0 y 1, donde un coeficiente de $GINI$ igual a 1 indica que el modelo genera una discriminación perfecta cuando el punto de corte varía en el intervalo [0,1].

El coeficiente de $GINI$ se relaciona con el $AUROC$ mediante la siguiente igualdad:

$$GINI = 2\,AUROC - 1 \tag{2.23}$$

de donde tenemos que el coeficiente de $GINI$ es 2 veces el área comprendida entre la curva ROC y la recta $y = x$

La forma usual de calcularlo se describe a continuación. Para analizar la discriminación a largo del intervalo [0,1], calculamos los deciles de la probabilidad pronosticada, en nuestro caso la probabilidad de activación P_A, y realizamos el siguiente cálculo:

$$GINI = 1 - \sum_{i=2}^{I} [N(i) + N(i-1)][P(i) - P(i-1)] \tag{2.24}$$

donde:

I: Número de intervalos, en este caso 10 intervalos (deciles).
N(i): Porcentaje acumulado de negativos hasta el intervalo i.
P(i): Porcentaje acumulado de positivos hasta el intervalo i.

Capítulo 3

Metodología analítica.

En este capítulo nos centraremos en describir la metodología empleada en la construcción del modelo de activación, para lo cual iniciaremos describiendo la población utilizada, definiremos la variable dependiente del modelo, construiremos y seleccionaremos las variables explicativas, ajustaremos un modelo de regresión logística a nuestra base de datos, para finalmente presentar los resultados y proceder a validar el modelo obtenido.

3.1. Construcción del modelo de activación.

3.1.1. Muestra de validación y modelamiento.

La muestra utilizada para el desarrollo del modelo se generó como sigue, primeramente se seccionó la población formada por todos los individuos que aperturaron una tarjeta de crédito en los meses enero 2012, abril 2012, julio 2012 y octubre 2012, los cuales serán denominados puntos de observación. En la Tabla 3.1, se presenta la distribución de los individuos de acuerdo al mes de apertura de la tarjeta de crédito.

Mediante el empleo de un muestreo aleatorio simple se seleccionó una muestra representativa de 22.803 individuos (60,50 % de la población total), la cual se dividió en dos submuestras aleatorias, con el objetivo de desarrollar el modelo con la primera (Muestra de modelamiento), y validarlo con la segunda (Muestra de validación).

Pto obs	Sujetos	Porcentaje	Acumulado
ene-12	4.888	13,4 %	13,4 %
abr-12	7.697	21,1 %	34,5 %
jul-12	10.652	29,2 %	63,7 %
oct-12	13.243	36,3 %	100,0 %
Total	36.480	100,0 %	

Tabla 3.1: Población total.

La submuestra que emplearemos en la validación del modelo corresponde al 20 % de la muestra original, es decir consta de 3.744 individuos, su distribución, de acuerdo al punto de observación, se presenta en la Tabla 3.2.

Pto obs	Sujetos	Porcentaje	Acumulado
ene-12	335	8,9 %	8,9 %
abr-12	552	14,7 %	23,6 %
jul-12	1.067	28,5 %	52,1 %
oct-12	1.790	47,9 %	100,0 %
Total	3.744	100,0 %	

Tabla 3.2: Muestra de validación.

En cambio la submuestra empleada en la construcción del modelo corresponde al 80 % de la muestra original, es decir consta de 19.059 individuos, su distribución de acuerdo al punto de observación se presenta en la Tabla 3.3.

Pto obs	Sujetos	Porcentaje	Acumulado
ene-12	1.796	9,4 %	9,4 %
abr-12	2.750	14,4 %	23,9 %
jul-12	5.313	27,9 %	51,7 %
oct-12	9.200	48,3 %	100,0 %
Total	19.059	100,0 %	

Tabla 3.3: Muestra de modelamiento.

Para seleccionar los individuos de la muestra final con la cual se va a construir el modelo es necesario realizar sobre ellos las siguientes consideraciones adicionales:

1. Registrar al menos una operación de crédito en el sistema crediticio ecuatoriano (Sistema financiero regulado, sistema regulado por la SEPS, sistema comercial) durante los últimos 36 meses anteriores al punto de observación.

2. Registrar al menos 3 meses de información, es decir registrar una antigüedad (tiempo transcurrido desde el primer crédito en meses) mayor a 3 meses en el sistema crediticio ecuatoriano.

con el propósito de disponer suficiente información histórica para poder predecir con mayor exactitud el comportamiento en la activación o no activación de la tarjeta dada la apertura, aquellos sujetos que no verifiquen alguna de estas condiciones no podrán ser evaluados con el modelo de activación que posteriormente presentaremos y deben ser catalogados como sin información. Después de realizar las consideraciones planteadas anteriormente obtenemos la muestra final de modelamiento cuya distribución de acuerdo al punto de observación se presenta en la Tabla 3.4. Los sujetos que verifican las condiciones 1 y 2 los denominaremos **bancarizados**.

Pto obs	Sujetos	Porcentaje	Acumulado
ene-12	1.211	7,2 %	7,2 %
abr-12	2.292	13,6 %	20,7 %
jul-12	4.755	28,1 %	48,9 %
oct-12	8.634	51,1 %	100,0 %
Total	16.892	100,0 %	

Tabla 3.4: Muestra de modelamiento final.

3.1.2. Generación de información de comportamiento y desempeño.

Para explicar de una manera didáctica la forma de generación de la información necesaria para desarrollar el modelo de activación consideremos la Figura 3.1. Los meses anteriores al punto de observación constituyen lo que denominaremos ventana de comportamiento, en el Ecuador por disposición de la SBS la ventana de comportamiento no puede ser superior a 36 meses, en este periodo se generan las variables asociadas al historial crediticio del individuo las cuales permiten evaluar su hábito de consumo, entre ellas tenemos deuda en tarjetas, cantidades de operaciones

de tarjeta, número de tarjetas con consumo y vigentes, cupos de tarjetas, montos de consumos, etc.

Las variables socio-demográficas tales como la edad, estado civil, provincia, region, etc. se generan al punto de observación.

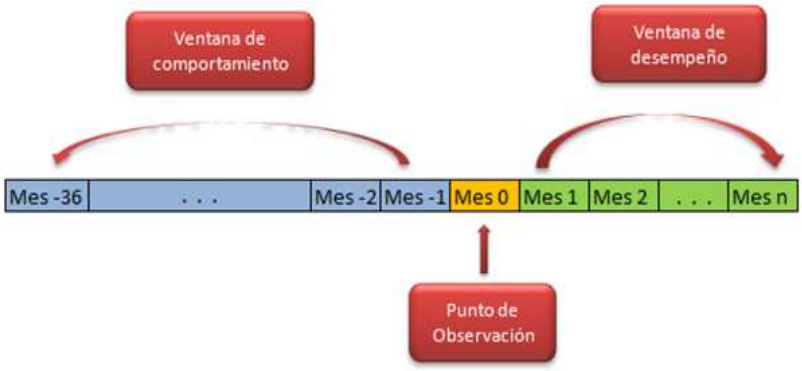

Figura 3.1: Generación de información.

Finalmente en los meses posteriores al punto de observación constituyen lo que denominaremos ventana de desempeño, generalmente de 12 o 24 meses evaluamos la conducta de consumo del individuo. La información generada en esta ventana es la que se utiliza para definir a los individuos Activa y a los No Activa (variable dependiente Y). El trabajo de [Nie et al., 2011], en el cual se estima la probabilidad de que un sujeto deje de realizar consumos con la tarjeta de crédito (deserción), permitió tener una idea del tipo de variables que se deberían generar para predecir la activación, puesto que la activación podría considerarse como el problema opuesto de la deserción.

3.1.3. Definición de la variable dependiente.

La variable dependiente Y será una variable binaria que toma el valor de 1 para los individuos etiquetados como Activa y 0 para los No activa. Emplearemos dos definiciones de activación, la primera que proponemos nosotros en el presente trabajo considera el tiempo transcurrido desde la apertura hasta la realización del primer consumo, y la segunda que considera la frecuencia y el tiempo entre consumos en la ventana de desempeño, esta definición fue propuesta en el trabajo realizado

por [Buckinx and Van den Poel, 2003].

Exigiremos adicionalmente que el monto de consumo sea mayor que una cierta cantidad fija $Consumo_{min}$ suficiente para cubrir al menos los gastos operativos de la apertura, el mantenimiento de la tarjeta y generar cierta rentabilidad para la institución, caso contrario, los sujetos serán etiquetados como indeterminados y no serán considerados en la construcción del modelo.

Compararemos las dos definiciones mencionadas y en este caso elegiremos aquella que mejor se pueda explicar con la información histórica disponible.

- **Definición 1.** Definimos la variable dependiente Y considerando el tiempo transcurrido desde que se aperturó la tarjeta hasta que se realizó el primer consumo en una ventana de desempeño de 12 meses. Con esta definición se busca etiquetar como Activa a aquellos sujetos que realizan el primer consumo tempranamente luego de la apertura de la tarjeta.

  Para definir la variable dependiente disponemos del saldo total de consumo para cada uno de los 12 meses de la ventana de desempeño incluyendo el punto de observación $saldo_0, saldo_1, ..., saldo_{12}$, consideraremos que un sujeto realiza al menos un consumo en el mes i, si $saldo_i > 0$.

  En la Tabla 3.5, presentamos la distribución de individuos respecto al mes en el cual se realiza el primer consumo. La columna Variación es la que emplearemos para la definición de Y, su forma de cálculo está dada por la expresión (3.1):

$$Variacion_i = \frac{Acumulado_i - Acumulado_{i-1}}{Acumulado_{i-1}} \% \quad i = 1, ..., 12. \tag{3.1}$$

  Básicamente analizamos la variación porcentual del número de individuos que realizan el primer consumo hasta el mes $i+1$ respecto del número de individuos que realizan el primer consumo hasta el mes i. Considerando la Figura 3.2, observamos que a partir del mes 4 la variación es aproximadamente nula, es decir que el mayor porcentaje de individuos realiza su primer consumo en

Mes 1^{er} Consumo	Sujetos	Porcentaje	Acumulado	Variación
0	3.907	23,1 %	23,1 %	
1	3.685	21,8 %	44,9 %	94,3 %
2	1.570	9,3 %	54,2 %	20,6 %
3	**912**	**5,4** %	**59,6** %	**9,9** %
4	464	2,7 %	62,4 %	4,5 %
5	313	1,9 %	64,2 %	2,9 %
6	266	1,6 %	65,8 %	2,4 %
7	173	1,0 %	66,8 %	1,5 %
8	127	0,8 %	67,6 %	1,1 %
9	138	0,8 %	68,4 %	1,2 %
10	115	0,7 %	69,1 %	0,9 %
11	112	0,7 %	69,7 %	0,9 %
12	79	0,5 %	70,2 %	0,6 %
No consume	**5.031**	**29,8** %	**100,0** %	
Total	16.892	100,0 %		

Tabla 3.5: Distribución por mes del primer consumo.

los tres primeros meses posteriores a la apertura, por lo tanto considerando como una variación significativa superior al 0,05 podemos definir la variable dependiente como sigue:

$$Y = \begin{cases} 1 & \text{Si el primer consumo se hace en los 3 meses posteriores a la apertura} \\ 0 & \text{Si no se realiza consumos en los 12 meses posteriores a la apertura} \end{cases} \tag{3.2}$$

Adicionalmente etiquetamos como Indeterminados a aquellos individuos que realizan el primer consumo entre los meses 4 y 12 después de la apertura o que realizan la activación con un monto menor a 40 dólares[1]. Estos individuos se excluyen del proceso de construcción del modelo con el propósito de aumentar la discriminación entre lo que se considera como Activa y como No Activa.

Finalmente en la Tabla 3.6, presentamos la distribución de la muestra de modelamiento de acuerdo a las categorías de la variable Y, podemos notar que

[1]El valor de 40 dólares se fijó tras consultar a la institución el monto monetario de activación suficiente para cubrir gastos operativos de la apertura y mantenimiento de la tarjeta y además generar una cierta rentabilidad para la entidad

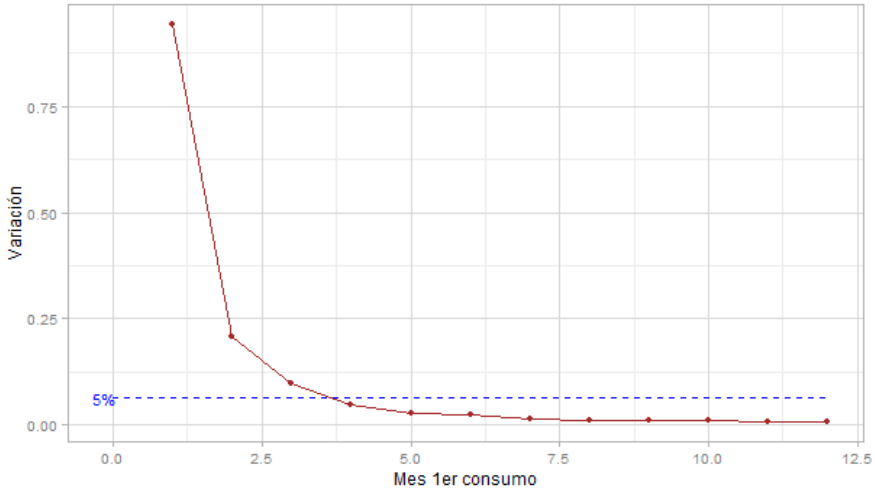

Figura 3.2: Mes del primer consumo

el porcentaje de sujetos Activa es mayor que el porcentaje de No Activa lo cual es lógico, además se debe considerar que el porcentaje de Indeterminados es relativamente bajo.

Y	Sujetos	Porcentaje	Acumulado
No Activa	5.031	29,8 %	29,8 %
Activa	10.074	59,6 %	89,4 %
Indeterminado	1.787	10,6 %	100,0 %
Total	16.892	100,0 %	

Tabla 3.6: Variable dependiente por mes del primer consumo.

- **Definición 2.** En este caso definimos la variable dependiente considerando la frecuencia de consumos y el tiempo entre consumos (tiempo inter-consumo) en una ventana de desempeño de 12 meses. Principalmente en esta definición se busca etiquetar a un sujeto como Activa si el número de consumos es superior al promedio y el tiempo inter-consumo es regular.

  Matemáticamente podemos definir el valor de la variable Y en el individuo i (y_i), como sigue, sean n el número de individuos de la muestra, C_i y Tic_i el número de consumos y el tiempo inter-consumo del individuo i respectivamente; $\overline{C}$ el promedio del número de consumos de la muestra. Notando como $\overline{Tic_i}$ al promedio del tiempo inter-consumo y como $\sigma^2_{Tic_i}$ a la varianza del tiempo inter-

consumo del individuo i en la ventana de desempeño podemos establecer la siguiente definición para la variable dependiente.

$$y_i = \begin{cases} 1 & \text{Si } C_i \geq \overline{C} \; y \; cv_i \leq \overline{cv}. \\ 0 & \text{Si no se realiza consumos en los 12 meses posteriores a la apertura.} \end{cases} \tag{3.3}$$

donde $cv_i = \frac{\overline{Tic_i}}{\overline{\sigma}_{Tic_i}}$, el cual se conoce como coeficiente de variación, es utilizado como una medida de la dispersión relativa, respecto a la media aritmética de una variable, se expresa en términos porcentuales.

Aquellos individuos que verifican que $C_i < \overline{C} \; y \; cv_i > \overline{cv}$ serán etiquetados como Indeterminados, aplicando esta definición a nuestra información tenemos la siguiente distribución de individuos de acuerdo a las categorías de la variable Y.

Y	**Sujetos**	**Porcentaje**	**Acumulado**
No Activa	4.616	27,3 %	27,3 %
Activa	8.827	52,3 %	79,6 %
Indeterminado	3.449	20,4 %	100,0 %
Total	16.892	100,0 %	

Tabla 3.7: Variable dependiente por frecuencia y tiempo entre consumos.

Analizando la Tabla 3.7, identificamos el inconveniente de que esta definición considerando nuestra información, genera un alto porcentaje de individuos que son Indeterminados.

3.1.4. Elección de la definición de la variable dependiente.

El estadístico de Kolomogorov Smirnov es uno de los estadísticos más utilizados para medir el poder predictivo de un modelo de regresión logística binaria, tal como se explicó en la sección (2.1.1) del capítulo 2.

Considerando la forma monótona creciente de la función logística tenemos que si una variable X presenta un KS de $q\,\%$, el modelo de regresión que incluya la variable X

presentará al menos un KS de $q\,\%$.

Para la definición 1 de la variable dependiente, en la Tabla 3.8, tenemos que la variable $Copen\_Vig\_3M$[2] presenta un $KS = 0,3685$, por lo tanto si generamos un modelo de regresión que incluya esta variable nuestro modelo presentará un KS de al menos $0,3685$.

i	**Variable**	**KS**
1	*Copen_Vig_3M*	0,3685
2	*TC_Abiert_Ult_3M*	0,3374
3	*Porc_TCcons_TCvig*	0,2677
4	*Tot_Cupo_TC*	0,2525
5	*Porc_Cupo_Util*	0,2447
6	*Tot_Saldo_Actual_TC*	0,1825
8	*Num_Acreedores_SFR*	0,1751
7	*Cred_TC_Sobre_Tot_Cred*	0,1622
9	*Max_Cupo_Vig_TC*	0,1349
10	*Max_Antiguedad_TC*	0,1308

Tabla 3.8: **10** principales variables por KS (Definición variable dependiente 1).

En cambio, si consideramos la definición 2 de la variable dependiente, en la Tabla 3.9, observamos que la variable $Copen\_Vig\_3M$ presenta un $KS = 0,2935$, es decir que si generamos un modelo de regresión que incluya esta variable nuestro modelo presentará un KS de al menos $0,2935$.

Por lo tanto con la información disponible obtendremos un modelo con un mayor poder predictivo si consideramos la variable dependiente Y definida en base al mes en el cual se realizó el primer consumo; comparando el KS de las variables restantes, observando las tablas Tabla 3.8 y Tabla 3.9, podemos decir que nuestra información permite explicar en mayor medida la definición 1 de la variable dependiente Y.

Por lo explicado anteriormente y considerando el hecho de que la definición 2 de la variable genera un gran porcentaje de individuos Indeterminados, se decide

[2]El significado de la notación utilizada en los nombres de todas las variables que emplearemos a lo largo del estudio se explica a detalle en el ANEXO A.

i	**Variable**	**KS**
1	*Copen_Vig_3M*	0,2935
2	*TC_Abiert_Ult_3M*	0,2672
3	*Tot_Cupo_TC*	0,2641
4	*Num_Acreedores_SFR*	0,2053
5	*Porc_Cupo_Util*	0,1994
6	*Porc_TCcons_TCvig*	0,1969
7	*Num_TC_Consumo*	0,1710
8	*Max_Cupo_Vig_TC*	0,1526
9	*Tot_Saldo_Actual_TC*	0,1520
10	*Cred_TC_Sobre_Tot_Cred*	0,1353

Tabla 3.9: **10** principales variables por KS (Definición variable dependiente 2).

considerar la **definición** 1, es decir el presente estudio se centrará en explicar la variable dada por la ecuación (3.2).

3.1.5. Filtrado de variables explicativas.

En esta sección seleccionaremos un subconjunto de variables explicativas, consideraremos aquellas que presenten la mayor divergencia entre las distribuciones de individuos Activa y No Activa, para ello calcularemos varias medidas de divergencia de distribuciones y estadísticos de pruebas de bondad de ajuste no paramétricas.

3.1.5.1. Filtrado de variables numéricas.

Para una variable numérica continua X_i dada, seleccionaremos los valores correspondientes a los individuos etiquetados como Activa y formaremos una nueva variable notada X_A y con los valores de la variable X_i de los individuos No Activa formaremos la variable X_{NA}. Nuestro objetivo es comparar las distribuciones de X_A y X_{NA} y así seleccionar las variables que presenten la mayor divergencia, estas variables permitirán obtener una mejor partición del conjunto de individuos en Activa y No Activa, es decir permitirán explicar de mejor manera la variable dependiente Y, tal como se explicó al analizar las medidas de divergencia en la sección (2.1) del capítulo 2.

Los criterios que emplearemos para seleccionar las variables numéricas continuas son los siguientes:

1. Estadístico de Kolmogorov-Smirnov (KS).
2. Estadístico Anderson Darling (AD).
3. Coeficiente de correlación de Pearson ($CORR$).

Para las variables numéricas disponibles calculamos el estadístico KS, AD, y el coeficiente de correlación $CORR$[3] y definimos un indicador (IND_{NUM}) del poder predictivo de una variable numérica mediante la combinación convexa de KS, AD, $CORR$, es decir:

$$IND_{NUM} = \alpha_1 KS + \alpha_2 AD + \alpha_3 |CORR| \tag{3.4}$$

con $\alpha_i > 0$, $i = 1, 2, 3$ y $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Los coeficientes $\alpha_1, \alpha_2, \alpha_3$ se pueden definir considerando la importancia que se quiera asignar a cada medida cuando evaluamos la discriminación del modelo final, en nuestro caso tomaremos $\alpha_1 = 0,60, \alpha_2 = 0,30, \alpha_3 = 0,10$, pues interesa obtener un modelo con el mayor KS posible, es por esta razón que asignamos el mayor peso $(0,60)$ a este estadístico.

Inicialmente se construyeron 50 variables numéricas, las cuales mediante el criterio experto se podrían considerar como aquellas que mejor expliquen la variable dependiente. En la Figura 3.3, presentamos el gráfico de sedimentación del número de variables y el indicador IND_{NUM}. Mediante este diagrama podemos decidir el número de variables a seleccionar considerando el correspondiente valor de IND_{NUM}.

Podemos observar que a partir de la variable 20, el indicador IND_{NUM} tiende a ser cero, es por esta razón que elegimos seleccionar únicamente las primeras 20 variables numéricas para analizar la posibilidad de incluirlas en el modelo de regresión o utilizarlas en la construcción de variables dummys y probabilidades de activa mediante la técnica de los árboles de decisión.

[3]El listado de todas las variables numéricas con los estadísticos KS, AD, coeficiente de correlación ($CORR$) se presenta en el ANEXO B.

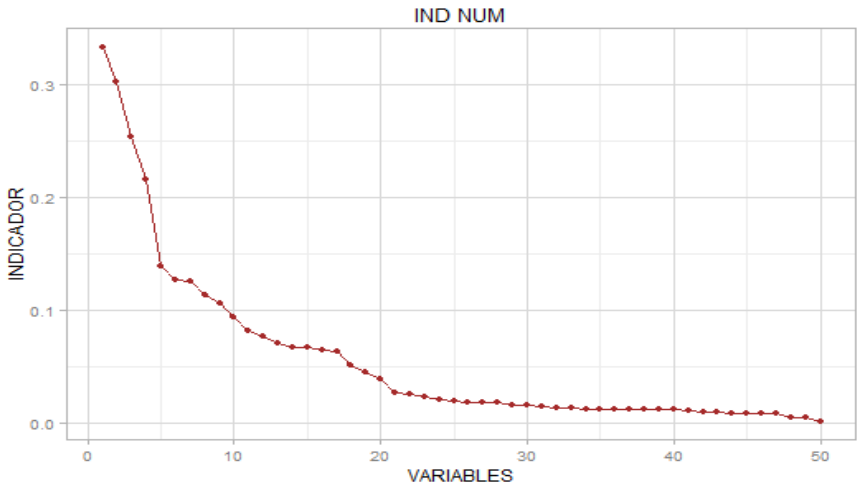

Figura 3.3: Gráfico de sedimentación variables numéricas.

3.1.5.2. Filtrado de variables categóricas.

Consideremos ahora una variable categórica Z_i con m categorías, para este tipo de variables se requiere que la información proporcionada por cada categoría sea diferente es decir que el número de individuos Activa en una categoría sea totalmente diferente al del resto de categorías, para medir esta diferencia utilizaremos los siguientes coeficientes, cuya interpretación y forma de cálculo se explicó en la sección (2.2) del capítulo 2.

1. Coeficiente de contingencia de Pearson ($CONT$).
2. Índice de valor de información (VI).
3. Diferencia de información por categoría (DIC)

Al igual que en las variables numéricas, para cada variable categórica calculamos $CONT$, VI, DIC[4] y formamos un indicador (IND_{CAT}) del poder predictivo de una variable categórica como la combinación convexa de los tres índices anteriores.

$$IND_{CAT} = \beta_1 CONT + \beta_2 VI + \beta_3 DIC \tag{3.5}$$

con $\beta_i > 0$, $i = 1,2,3$ y $\beta_1 + \beta_2 + \beta_3 = 1$. Los coeficientes $\beta_1, \beta_2, \beta_3$ se pueden definir considerando la importancia que se le quiera asignar a cada medida cuando

[4]El listado de todas las variables categóricas con los índices $CONT$, VI, DIC se presenta en el ANEXO C.

evaluamos la discriminación del modelo final, en nuestro caso tomaremos $\beta_1 = 0,60, \beta_2 = 0,30, \beta_3 = 0,10$, de tal forma de asignar el mayor peso $(0,60)$ al estadístico que mide la dependencia de las variables.

En la Figura 3.4, se presenta el gráfico de sedimentación para el indicador IND_{CAT}, en este caso disponemos de cantidad baja de variables categóricas, de acuerdo al diagrama elegimos únicamente 10 variables de las 12 disponibles, éstas serán las que utilizaremos para construir las variables dummys y probabilidades de activa mediante la técnica de los árboles de decisión.

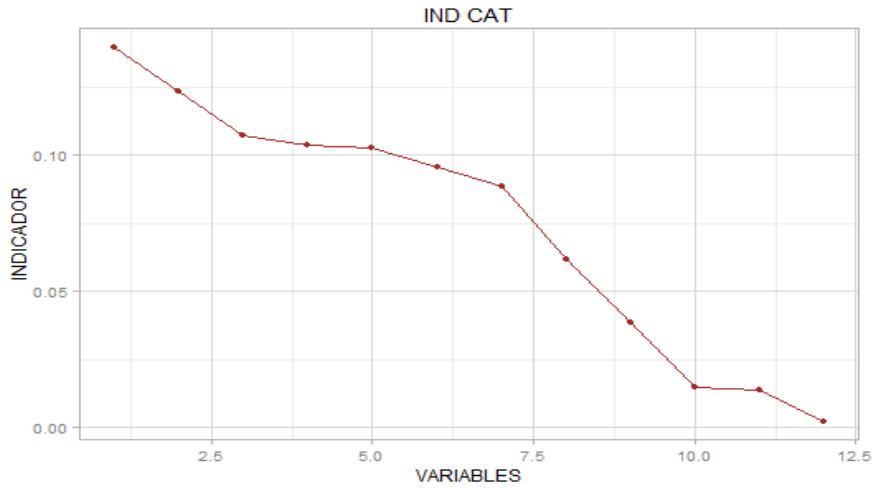

Figura 3.4: Gráfico de sedimentación variables categóricas.

Mediante la utilización de las medidas anteriores obtenemos una nueva base de datos, la cual consta de las variables numéricas y categóricas con el mayor poder predictivo. Este filtrado lo realizamos con el propósito de descartar una gran cantidad de variables que no influyen en la discriminación de individuos en Activa y No Activa.

3.1.6. Generación de variables explicativas.

Antes de proceder a describir las variables explicativas utilizadas en el modelo de regresión con el propósito de predecir la probabilidad de activación, su forma de generación y la justificación de su utilización, presentamos un diagrama de flujo en el cual se detallan cada una de las etapas que se deben seguir para la construcción

de un modelo de predicción de una variable dependiente binaria.

En el flujograma de la Figura 3.5, se describen cada uno de los pasos necesarios para construir el modelo de activación, comenzando con la extracción de las variables dependiente y explicativas de una determinada base de datos para finalmente generar el modelo de regresión logística que estima la probabilidad de activación y los resultados necesarios para validarlo.

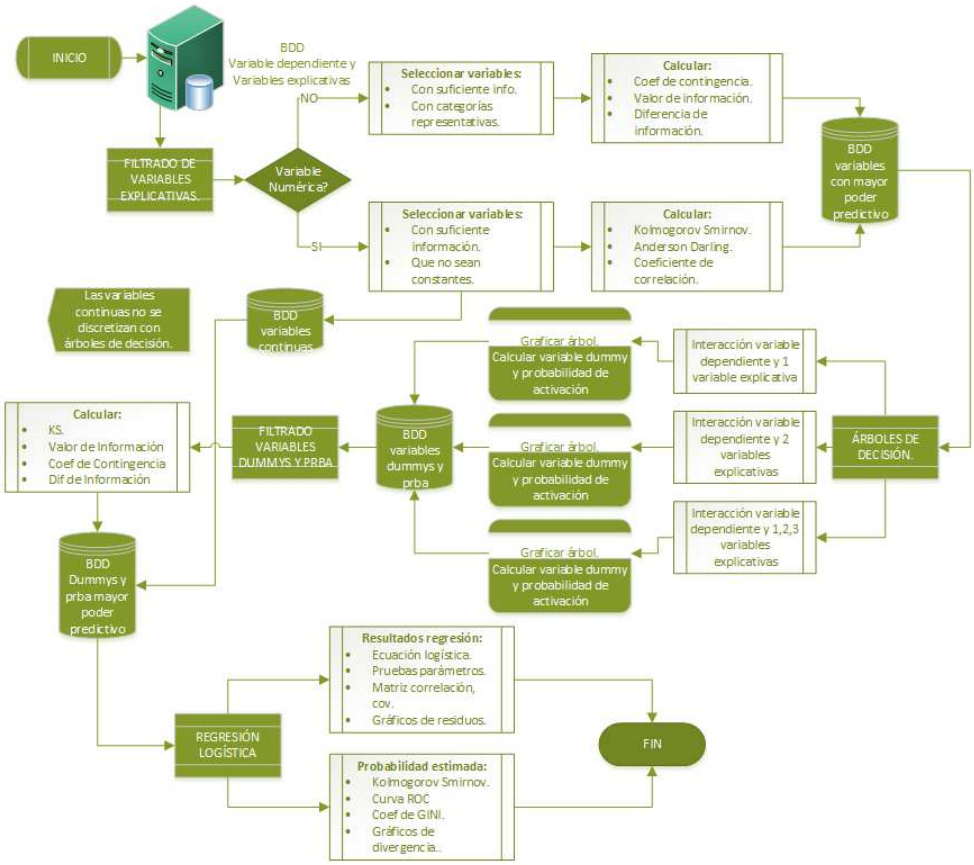

Figura 3.5: Gráfico de sedimentación variables categóricas.

Luego de probar varios modelos de regresión y en base a indicadores tales como el R^2, el porcentaje de clasificación correcta, el estadístico de Kolmogorov-Smirnov KS, el área bajo la curva ROC $(AUROC)$, el índice de $GINI$, se obtuvo el modelo constituido por las variables explicativas de la Tabla 3.10, el mismo que presenta un coeficiente de determinación $R^2 = 0,4245$; considerando el tipo de información disponible y el tipo de modelo construido podemos decir que el ajuste del modelo es aceptable.

En la Tabla 3.10 presentamos el nombre de la variable explicativa (Variable) con su respectivo parámetro estimado (Coeficiente), error estándar, significancia, y los extremos del intervalo de confianza de nivel 95 %, extremo inferior (Lim inf) y extremo superior (Lim sup).

i	Variable	Coeficiente	Error estándar	Significancia	Lim inf	Lim sup
1	*Porc_Cupo_Util*	0,009	0,001	0,000	0,007	0,012
2	*Max_Antiguedad_TC*	-0,007	0,001	0,000	-0,008	-0,006
3	*Ln_Max_Cupo_Vig_TC*	0,120	0,012	0,000	0,096	0,144
4	*Ln_Cupo_Prom_TC_Vig*	0,187	0,008	0,000	0,171	0,203
5	*Mejor_TC*	-0,505	0,037	0,000	-0,579	-0,432
6	*rdop_3ab12_TC*	0,238	0,117	0,041	0,008	0,467
7	*rdop_3ab12_C*	0,131	0,034	0,027	-0,005	0,267
8	*Ln_rfp24_C*	0,021	0,010	0,032	0,002	0,040
9	*Saldo_Cred_Comercial_12M*	-0,172	0,061	0,005	-0,292	-0,051
10	*Ln_Saldo_Sicom*	0,021	0,005	0,000	0,012	0,031
11	*Comprometido*	-0,420	0,113	0,000	-0,641	-0,199
12	*d_Azuay*	0,315	0,085	0,000	0,148	0,482
13	*prba_Region*	3,353	0,465	0,000	2,442	4,264
14	*prba_Edad*	1,335	0,530	0,012	0,296	2,373
15	*prba_TC_Abiert_Ult_3M y Copen_Vig_3M*	4,461	0,105	0,000	4,256	4,666
16	Constante	-2,452	0,587	0,000	-3,603	-1,301

Tabla 3.10: Variables explicativas en el modelo de regresión logística.

A continuación se describen a detalle las variables de la Tabla 3.10.

1. **Porc_Cupo_Util:** Porcentaje utilizado del cupo total que registra el sujeto al punto de observación.

$$Porc\_Cupo\_Util = \frac{Total\_Saldo\_TC}{Total\_Cupo\_TC}$$

donde:

- $Total\_Saldo\_TC$: Suma de los saldos de todas las tarjetas de crédito que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.
- $Total\_Cupo\_TC$: Suma de los cupos de todas las tarjetas de crédito que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.

La variable $Porc\_Cupo\_Util$ es significativa y tiene signo positivo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea el porcentaje de cupo utilizado, mayor será la probabilidad de activación, esto resulta lógico pues implicaría que el cliente mantiene activas sus tarjetas de crédito.

2. **Max_Antiguedad_TC:** Tiempo transcurrido en meses desde que el sujeto aperturó su primera tarjeta de crédito hasta el punto de observación en cualquiera de las instituciones pertenecientes al sistema financiero regulado por la SBS.

   La variable $Max\_Antiguedad\_TC$ es significativa y tiene signo negativo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea la antigüedad en TC menor será la probabilidad de activación, esto resulta lógico pues un cliente que aperturó su primera tarjeta hace 6 meses es más propenso a activar una nueva tarjeta que uno que haya aperturado una tarjeta hace 24 meses.

3. **Ln_Max_Cupo_Vig_TC:** Logaritmo natural del máximo cupo de las tarjetas de crédito vigentes, es decir que no hayan sido canceladas, que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.

$$Ln\_Max\_Cupo\_Vig\_TC = \begin{cases} \ln(Max\_Cupo\_TC) & \text{Si } Max\_Cupo\_TC > 0 \\ \ln(k) & \text{Si } Max\_Cupo\_TC = 0 \end{cases}$$

   donde:

- Max_Cupo_TC: Máximo cupo de las tarjetas de crédito vigentes que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.

Se debe considerar que pueden existir ciertos sujetos para los cuales la variable Max_Cupo_TC sea igual a 0, en este caso la variable Ln_Max_Cupo_Vig_TC no estaría bien definida, es por esta razón que para los sujetos que registren un $Max\_Cupo\_TC = 0$, les asignamos un valor positivo k lo suficientemente pequeño, por ejemplo 0,01 dólares.

4. **Ln_Cupo_Prom_TC_Vig:** Logaritmo natural del cupo promedio por tarjeta de crédito vigente.

$$Cupo\_Prom\_TC = \begin{cases} \frac{Total\_Cupo\_TC}{Num\_TC\_Vig} & \text{Si } Num\_TC\_Vig > 0 \\ 0 & \text{Si } Num\_TC\_Vig = 0 \end{cases}$$

$$Ln\_Cupo\_Prom\_TC\_Vig = \begin{cases} \ln(Cupo\_Prom\_TC) & \text{Si } Cupo\_Prom\_TC > 0 \\ \ln(k) & \text{Si } Cupo\_Prom\_TC = 0 \end{cases}$$

donde:

- $Total\_Cupo\_TC$: Suma de los cupos de todas las tarjetas que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.
- Num_TC_Vig: Número de tarjetas vigentes que posee el sujeto al punto de observación en las instituciones pertenecientes al sistema financiero regulado por la SBS.

Análogamente que para la variable Max_Cupo_TC, asignamos un valor positivo k lo suficientemente pequeño, por ejemplo 0,01 dólares a la variable Cupo_Prom_TC cuando esta es igual a 0.

Tanto la variable $Ln\_Max\_Cupo\_Vig\_TC$ como $Ln\_Cupo\_Prom\_TC\_Vig$ resultan significativas y tienen signo positivo en el modelo de regresión logística

(ver Tabla 3.10), lo cual quiere decir que mientras mayor sea cupo máximo o el cupo promedio del sujeto, mayor será la probabilidad de activación, esto resulta lógico pues se asignan cupos altos para aquellos clientes con un buen hábito de consumo, los cuales serían más propensos a activar la TC.

La transformación logarítmica se utiliza para obtener un modelo estable en el tiempo, puesto que si el cupo de un sujeto en el mes i es de 500 dólares y en el mes $i+1$ es elevado a 4.000 dólares, el impacto sobre la probabilidad de activación sería muy elevado. Para atenuar este efecto se acostumbra a utilizar la transformación logarítmica.

5. **Mejor_TC:** La variable $Mejor\_TC$ es una variable numérica construida utilizando el criterio del conocimiento del negocio, consiste en tomar el máximo de las codificaciones de las tarjetas que posee un sujeto en las instituciones pertenecientes al sistema financiero regulado por la SBS, de acuerdo a su importancia, la cual está dada a través de la Tabla 3.11.

i	**Nombre de tarjeta**	**Código_TC**
1	DINERS	10
2	MASTERCARD	9
3	VISA	8
4	AMERICAN EXPRESS	7
5	DISCOVER	6
6	CUOTA FACIL - UNIBANCO	5
7	VISA CASH - BANCO DEL PACIFICO	4
8	ROSE-BANCO INTERNACIONAL	3
9	COOPCARD	2
10	FILANCARD	1
11	OTROS	1
12	Tarjetas de Sistemas cerrados	0
13	CREDITO SI - BANCO TERRITORIAL	0

Tabla 3.11: Codificación del tipo de tarjeta de crédito.

Por ejemplo si un sujeto posee las tarjetas MASTERCARD, VISA, DISCOVER, la variable $Mejor\_TC$ se obtendrá como sigue:

$$Mejor\_TC = Max\{9,8,6\} = 9$$

La variable $Mejor\_TC$ resulta significativa y tiene signo negativo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mejor sea la tarjeta que posee el sujeto, menor será la probabilidad de activación de una nueva TC, esto resulta lógico pues generalmente un cliente que posea las tarjetas DINERS y ROSE es menos propenso a activar la tarjeta ROSE la cual tiene cupos más bajos y su obtención no requiere de tantas restricciones (ingreso, score de comportamiento, etc).

6. **rdop_3ab12_TC:** Razón entre el número de operaciones de tarjeta de crédito realizadas en los últimos 3 meses respecto al número de operaciones de tarjetas realizadas en los últimos 12 meses anteriores al punto de observación.

$$rdop\_3ab12\_TC = \begin{cases} \dfrac{Copen\_Vig\_3\_TC}{Copen\_Vig\_12\_TC} & \text{Si } Copen\_Vig\_12\_TC > 0 \\ 0 & \text{Si } Copen\_Vig\_12\_TC = 0 \end{cases}$$

donde:

- $Copen\_Vig\_3\_TC$: Cantidad de operaciones de tarjeta de crédito realizadas durante los últimos 3 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.
- $Copen\_Vig\_12\_TC$: Cantidad de operaciones de tarjeta de crédito realizadas durante los últimos 12 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.

La variable $rdop\_3ab12\_TC$ resulta significativa y tiene signo positivo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea la razón entre las operaciones de TC realizadas en los 3 meses respecto a las realizadas en los 12 meses anteriores al punto de observación, mayor será la probabilidad de activación, esto resulta lógico pues si se incrementan las operaciones en los 3 meses anteriores al punto de observación tendríamos que el cliente presenta una variación en la frecuencia de consumo.

7. **rdop_3ab12_C:** Razón entre el número de operaciones de consumo realizadas durante los últimos 3 meses respecto al número de operaciones de consumo realizadas durante los últimos 12 anteriores al punto de observación.

$$rdop\_3ab12\_C = \begin{cases} \frac{Copen\_Vig\_3\_Consumo}{Copen\_Vig\_12\_Consumo} & \text{Si } Copen\_Vig\_12\_Consumo > 0 \\ 0 & \text{Si } Copen\_Vig\_12\_Consumo = 0 \end{cases}$$

donde:

- $Copen\_Vig\_3\_Consumo$: Cantidad de operaciones de consumo realizadas durante los últimos 3 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.
- $Copen\_Vig\_12\_Consumo$: Cantidad de operaciones de consumo realizadas durante los últimos 12 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.

La variable $rdop\_3ab12\_C$ resulta significativa y tiene signo positivo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea la razón entre las operaciones de consumo realizadas en los 3 meses respecto a las realizadas en los 12 meses anteriores al punto de observación, mayor será la probabilidad de activación, esto resulta lógico pues los créditos de consumo están muy relacionados con las operaciones de tarjeta de crédito.

Se analizó la necesidad de la inclusión de este tipo de variables, puesto que puede suceder que sea necesario evaluar a un individuo que no haya aperturado una tarjeta en los últimos 36 meses anteriores al punto de observación, este individuo no tendría información histórica referente a TC por ello sería necesario analizar su comportamiento en créditos similares, por ejemplo consumo.

8. **Ln_rfp24_C:** Logaritmo natural de la razón de la deuda por vencer en créditos de consumo respecto a la deuda total en créditos de consumo en los últimos 24 meses anteriores al punto de observación.

$$rfp24\_C = \begin{cases} \frac{Deuda\_por\_vencer\_24m\_Consumo}{Deuda\_Total\_24m\_Consumo} & \text{Si } Deuda\_Total\_24m\_Consumo > 0 \\ 0 & \text{Si } Deuda\_Total\_24m\_Consumo = 0 \end{cases}$$

$$Ln\_rfp24\_C = \begin{cases} \ln(rfp24\_C) & \text{Si } rfp24\_C > 0 \\ \ln(k) & \text{Si } rfp24\_C = 0 \end{cases}$$

donde:

- $Deuda\_por\_vencer\_24m\_Consumo$: Deuda total por vencer en créditos de consumo en los últimos 24 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.
- $Deuda\_Total\_24m\_Consumo$: Deuda total en créditos de consumo en los últimos 24 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.

Análogamente que para la variable Max_Cupo_TC, asignamos un valor positivo k lo suficientemente pequeño, por ejemplo 0,01 dólares a la variable rfp24_C cuando esta es igual a 0.

El análisis de esta variable es similar a la variable $rdop\_3ab12\_C$.

9. **Saldo_Cred_Comercial_12M:** Variable binaria que toma el valor de 1 si el máximo saldo en operaciones comerciales en entidades reguladas por la SBS durante en los últimos 12 meses anteriores al punto de observación es mayor que cero.

$$Saldo\_Cred\_Comercial\_12M = \begin{cases} 1 & \text{si} Max\_Saldo\_Ope\_Comercial > 0 \\ 0 & \text{si} Max\_Saldo\_Ope\_Comercial = 0 \end{cases}$$

donde:

- $Max\_Saldo\_Ope\_Comercial$: Máximo saldo en operaciones de crédito comercial vigentes durante los últimos 12 meses anteriores al punto de observación en las instituciones pertenecientes al Sistema Financiero Regulado por la SBS.

La variable $Saldo\_Cred\_Comercial\_12M$ resulta significativa y tiene signo negativo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere

decir que si el máximo saldo en operaciones comerciales en entidades reguladas por la SBS durante en los últimos 12 meses anteriores al punto de observación es mayor que cero, menor será la probabilidad de activación, esto se puede justificar porque generalmente los créditos comerciales son realizados por microempresarios y no por asalariados quienes generalmente son más propensos a la activación.

10. **Ln_Saldo_Sicom:** Logaritmo natural de la sumatoria del valor total de la deuda en el sector comercial al punto de observación.

$$Ln\_Saldo\_Sicom = \begin{cases} \ln(Saldo\_Sicom) & \text{Si } Saldo\_Sicom > 0 \\ \ln(k) & \text{Si } Saldo\_Sicom = 0 \end{cases}$$

donde:

- $Saldo\_Sicom$: Deuda total al punto de observación en las instituciones pertenecientes al sistema comercial.

Análogamente que para la variable Max_Cupo_TC, asignamos un valor positivo k lo suficientemente pequeño, por ejemplo 0,01 dólares a la variable Saldo_Sicom cuando esta es igual a 0.

La variable Ln_Saldo_Sicom resulta significativa y tiene signo positivo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea la sumatoria del valor total de la deuda en el sector comercial mayor será la probabilidad de activación, esto se puede justificar porque un sujeto realiza la mayoría de las operaciones de tarjeta en el sector comprendido por las casas comerciales.

11. **Comprometido:** Razón de la cuota estimada respecto al ingreso que registra el sujeto al punto de observación.

$$Comprometido = \frac{Cuota\_Estimada}{Ingreso}$$

donde:

- $Cuota\_Estimada$: La cuota estimada es una variable que corresponde a la cuota mensual que tiene que cancelar un sujeto con respecto al total del endeudamiento en el Sistema Crediticio Ecuatoriano, en donde tenemos el sistema regulado por la SBS, por la SEPS, y el sector correspondiente a las casas comerciales.
- $Ingreso$. Se puede utilizar el ingreso mensual real que registre el sujeto o en su defecto generar un modelo de regresión que permita estimarlo considerando la información disponible.

La variable $Comprometido$ es significativa y tiene signo negativo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que mientras mayor sea el porcentaje de endeudamiento en el Sistema Crediticio Ecuatoriano respecto al ingreso, menor será la probabilidad de activación.

12. **d_Azuay:** Variable binaria que toma el valor de 1 si el individuo pertenece a la provincia del AZUAY y 0 caso contrario, su cálculo se realiza a través de la extracción de los dos primeros dígitos de la identificación.

$$d\_Azuay = \begin{cases} 1 & \text{Si los dos primeros dígitos de la identificación son 01.} \\ 0 & \text{Si no.} \end{cases}$$

La variable d_Azuay es significativa y tiene signo positivo en el modelo de regresión logística (ver Tabla 3.10), lo cual quiere decir que un sujeto que pertenezca a la provincia del Azuay es más propenso a activar una nueva tarjeta que un cliente de cualquier otra provincia, esto generalmente se debe a la lealtad de los clientes con la institución, la cual tiene su sede en esta provincia.

Las siguientes variables que explicaremos fueron construidas mediante la utilización de la técnica de los árboles de decisión explicada a detalle en la sección (2.3) del capítulo 2.

13. **prba_Region:** Probabilidad de Activa de la variable Región construida mediante el árbol de decisión de la Figura 3.6, en el cual podemos observar que el porcentaje de sujetos Activa en el Nodo 1 ($72,36\,\%$) es mayor que el porcentaje

de sujetos Activa en el Nodo 2 ($62,36\,\%$), lo cual nos dice que un sujeto de la COSTA tiene mayor probabilidad de activación que un sujeto perteneciente a cualquier otra región.

$$prba\_Region = \begin{cases} 0,7236 & \text{Si Region = COSTA.} \\ 0,6236 & \text{Si Region} \neq \text{COSTA.} \end{cases}$$

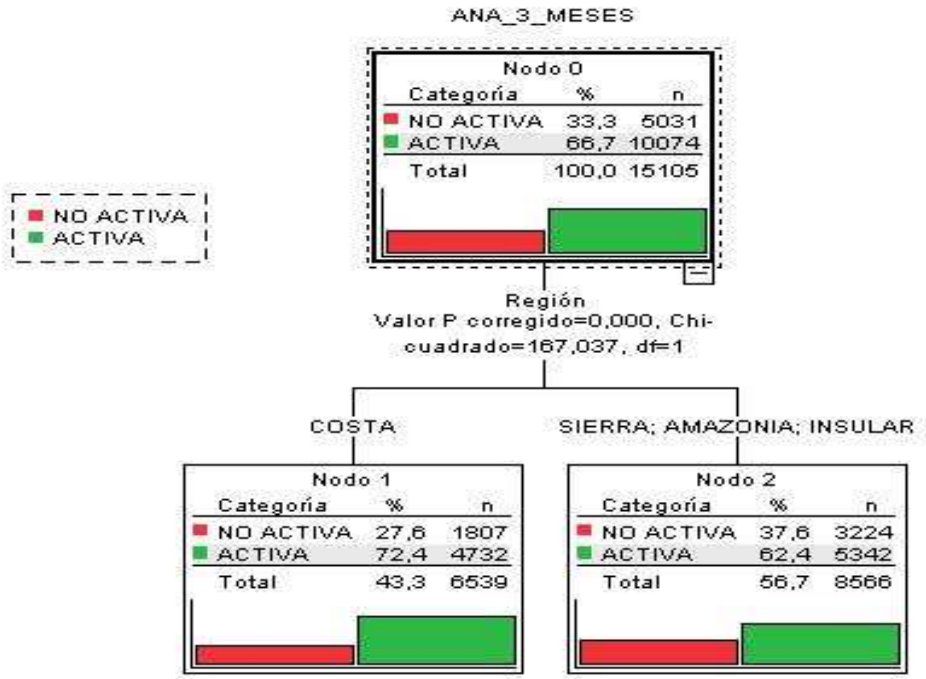

Figura 3.6: Árbol de decisión para la variable Región.

14. **prba_Edad:** Probabilidad de activación de la variable Edad construida mediante el árbol de decisión de la Figura 3.7, en el cual podemos observar que la probabilidad de activación disminuye con el incremento de la edad.

    Los sujetos con una edad menor o igual a 27 años (Nodo 1) tienen una probabilidad de activación de $0,7730$, los sujetos con una edad mayor a 27 y menor o igual que 35 años (Nodo 2) tienen una probabilidad de activación de $0,6830$, los sujetos con una edad mayor a 35 y menor o igual que 51 años (Nodo 3) tienen una probabilidad de activación de $0,6470$, finalmente los sujetos con una edad mayor a 51 años (Nodo 4) tienen una probabilidad de activación de

$0,6250$.

$$prba\_Edad = \begin{cases} 0,773 & \text{Si } Edad \leq 27. \\ 0,683 & \text{Si } 27 < Edad \leq 35. \\ 0,647 & \text{Si } 35 < Edad \leq 51. \\ 0,625 & \text{Si } Edad > 51. \end{cases}$$

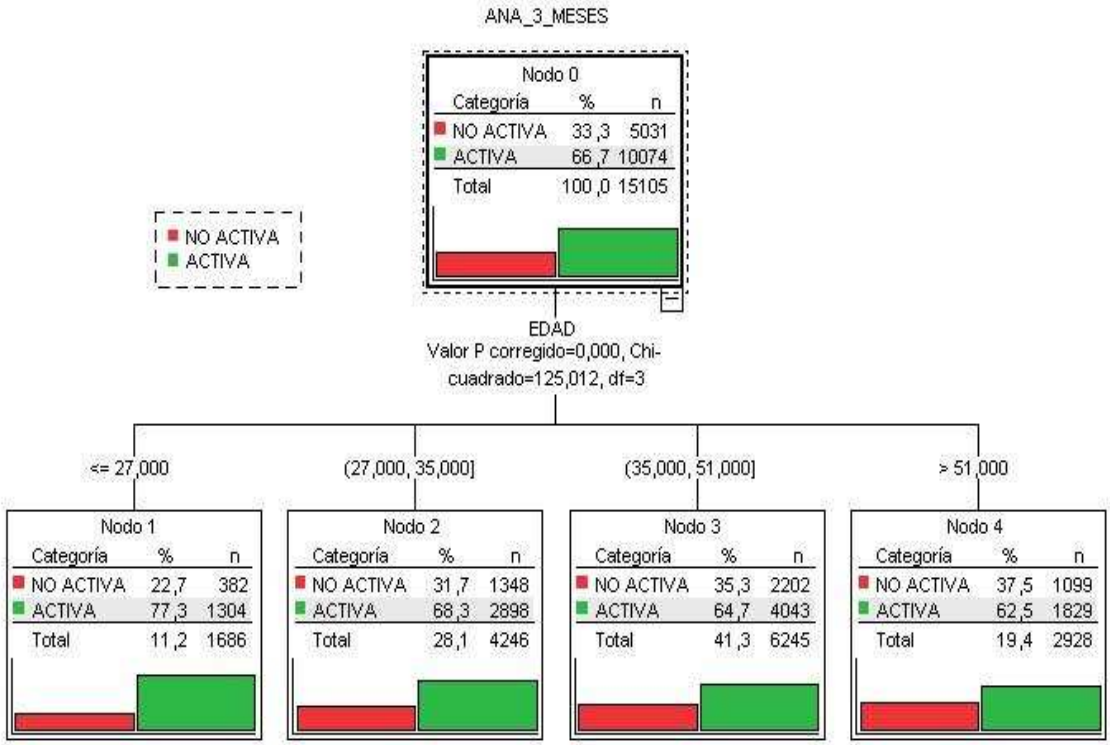

Figura 3.7: Árbol de decisión para la variable Edad.

15. **prba_TC_Abiert_Ult_3M y Copen_Vig_3M:** Esta es una variable construida a través de la interacción de las variables:

    a) $TC\_Abiert\_Ult\_3M$: Número de tarjetas abiertas durante los últimos 3 meses anteriores al punto de observación.

    b) $Copen\_Vig\_3M$: Variable binaria que toma el valor de 1 si el sujeto registra uno o más productos vigentes, abiertos durante los últimos 3 meses anteriores al punto de observación.

    Dicha construcción la podemos visualizar en el árbol de decisión de la Figura 3.8, matemáticamente tenemos la variable:

$$prba\_TC\_Abiert\_Ult\_3M\,y\,Copen\_Vig\_3M = \begin{cases} 0,415 & \text{Si } TC\_Abiert\_Ult\_3M \leq 1 \\ & \text{y } Copen\_Vig\_3M \leq 0 \\ 0,798 & \text{Si } TC\_Abiert\_Ult\_3M \leq 1 \\ & \text{y } Copen\_Vig\_3M > 0 \\ 0,884 & \text{Si } 1 < TC\_Abiert\_Ult\_3M \leq 2 \\ & \text{y } Copen\_Vig\_3M \leq 0 \\ 0,938 & \text{Si } 1 < TC\_Abiert\_Ult\_3M \leq 2 \\ & \text{y } Copen\_Vig\_3M > 0 \\ 0,697 & \text{Si } TC\_Abiert\_Ult\_3M > 2 \end{cases}$$

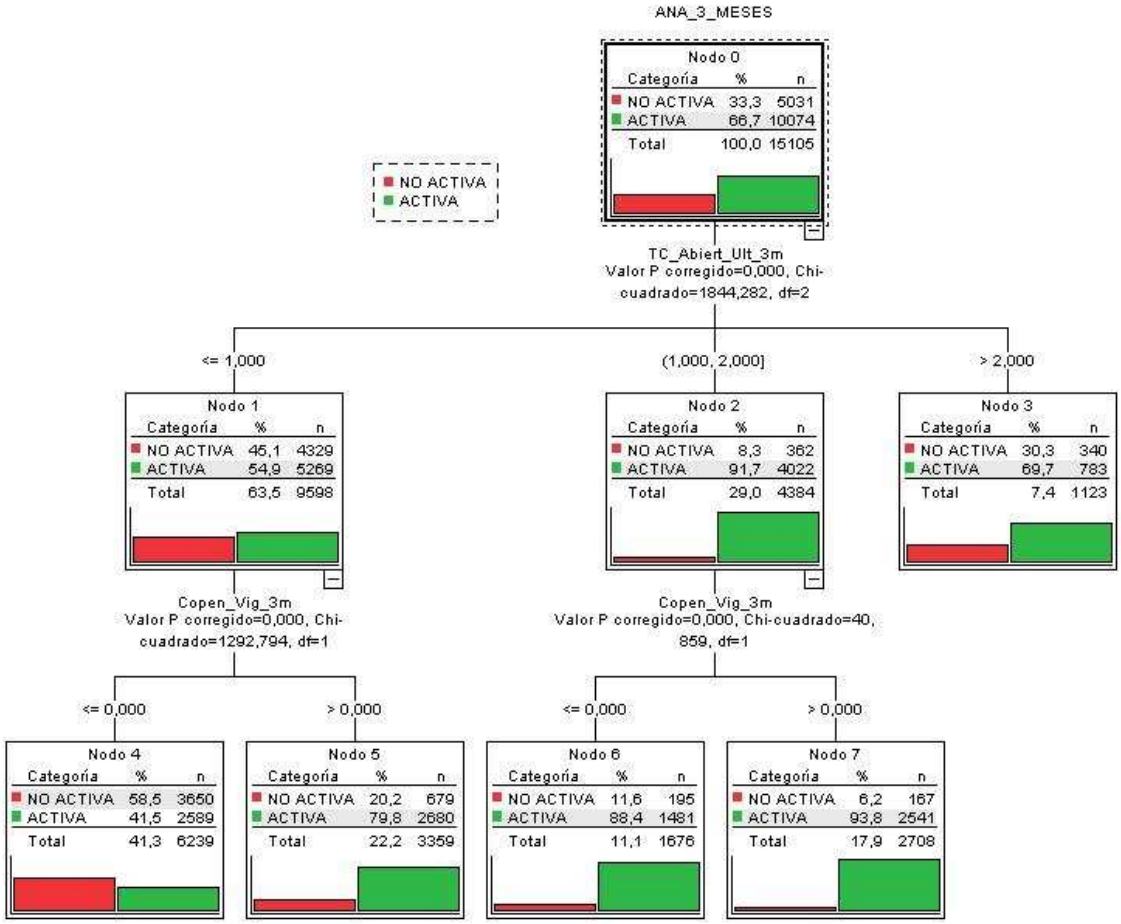

Figura 3.8: Árbol de decisión, variables TC_Abiert_Ult_3M y Copen_Vig_3M.

Esta interacción de variables se realiza con el objetivo de incrementar el poder discriminativo de las variables, medido a través del KS, en este caso el KS de la variable $TC\_Abiert\_Ult\_3M$ es de $0,3685$ y el de la variable $Copen\_Vig\_3M$ de

$0,3374$, tras la interacción mediante el árbol de decisión obtenemos la nueva variable $prba\_TC\_Abiert\_Ult\_3M\,y\,Copen\_Vig\_3M$, la cual presenta un KS de $0,498$.

3.1.7. Resultados y validación del modelo de regresión logística.

En esta sección nos centraremos en presentar varios resultados obtenidos a través de los cuales analizaremos la calidad de discriminación y predicción del modelo de regresión logística obtenido en la sección anterior.

1. **Multicolinealidad:** [Castro, 2008] define a la multicolinealidad como el problema de que una variable explicativa incluida en el modelo de regresión sea aproximadamente una combinación lineal de las restantes, es decir que los regresores estén fuertemente correlacionadas. Esto ocasiona que el determinate de la matriz X^tX, utilizada en la estimación de los parametros, sea cercano a 0, lo cual implica que su inversa no es exacta y la estimación de cada parámetro es inestable, es decir pequeñas variaciones en los decimales de los datos (variables explcativas) ocasionan enormes variaciones en el resultado (parámetros), es decir nos encontramos ante la presencia de un problema mal condicionado.

   Para estudiar el problema de multicolinealidad, utilizaremos el índice de condicionamiento (IC), el mismo que esta dado mediante la expresión

   $$IC = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

   donde: $\lambda_{max}, \lambda_{min}$, son los valores propios máximo y mínimo respectivamente, de la matriz de correlaciones de los regresosres. En nuestro caso $\lambda_{max} = 1,58$ y $\lambda_{min} = 0,36$, luego:

   $$IC = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{1,58}{0,36}} = 2,07$$

   [Castro, 2008] índica que si $IC < 10$, no hay presencia de multicolinealidad; si $10 \leq IC \leq 15$, existe una multicolinealidad moderada; y si $IC > 15$, existe una multicolinealidad fuerte. En nuestro caso $IC = 2,07 < 10$, lo que implica que no hay presencia de multicolinealidad.

2. **Medidas de calidad de discriminación.**

   - **KS:** En este numeral analizamos la máxima diferencia entre las distribuciones de acumulación empíricas de la probabilidad de activación estimada para sujetos Activa y No Activa, dada por el estadístico de Kolmogorov-Smirnov explicado a detalle en la sección (2.1.1) del capítulo 2. Las líneas de código para el cálculo del KS en R son las siguientes:

```
# Pr_A: Probabilidad de activación de sujetos Activa
# Pr_NA: Probabilidad de activación de sujetos No Activa
  ks.test(Pr_A,Pr_NA)


Two-sample Kolmogorov-Smirnov test
data:  x1 and x2
D = 0,5544, p-value < 2.2e-16
alternative hypothesis: two-sided
```

   Como vemos en el código anterior para el modelo de regresión logística final se obtuvo un KS del $0,5544$ considerando la muestra de modelamiento, este valor indica una alta divergencia entre las distribuciones de acumulación empíricas de Pr_A y Pr_{NA}. Teniendo en cuenta la información disponible y el tipo de modelo realizado este valor permite concluir que el modelo es altamente discriminativo.

   Para evitar el problema de sobreajuste del modelo a los datos de modelamiento, es decir que el modelo únicamente presente buena discriminación únicamente para la información con la cual fue desarrollado es necesario evaluarlo con una base distinta a la de modelamiento, en nuestro caso evaluamos el modelo utilizando **la muestra de validación** (descrita en la sección 3.1.1 del capítulo 3). Es por esta razón que de forma análoga que para la muestra de modelamiento presentamos ahora el KS del modelo considerando la muestra de validación, el valor para el estadístico que se obtuvo fue de $0,5209$, el cual es muy similar al obtenido utilizando la muestra de modelamiento, lo que nos permite decir que el modelo no está

sobreajustado a la información de la muestra de modelamiento.

En la Figura 3.9a tenemos la representación gráfica del KS utilizando la muestra de modelamiento, en esta figura podemos observar la divergencia entre las distribuciones de acumulación empíricas de Pr_A y Pr_{NA}, y en la Figura 3.9b en cambio presentamos el gráfico correspondiente a la muestra de validación, el cual es muy similar al obtenido utilizando la muestra de desarrollo.

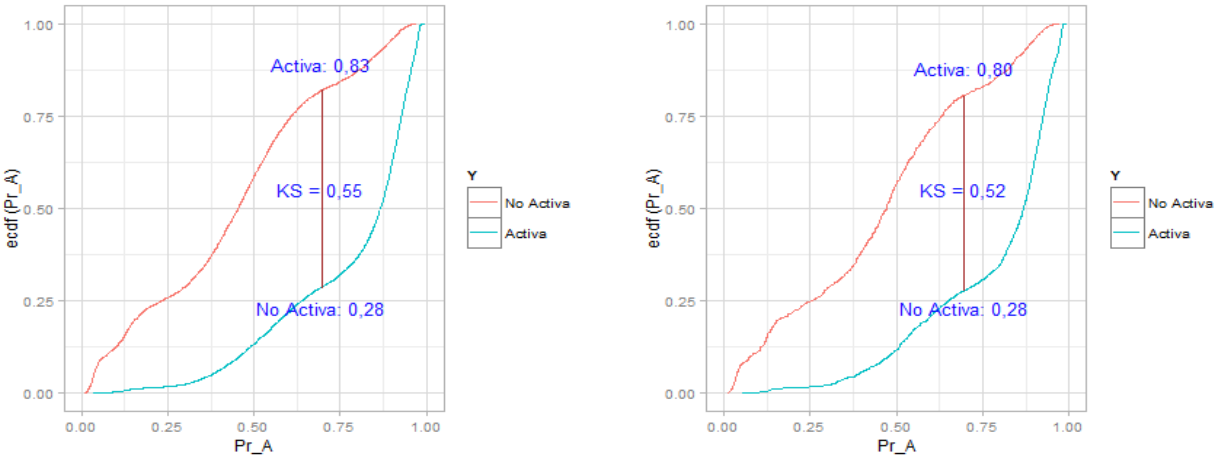

Figura 3.9: Curva KS muestra de modelamiento-validación.

- **Área bajo la curva ROC:** En la Figura 3.10a, tenemos la representación gráfica de la curva ROC, con la probabilidad de activación estimada utilizando la muestra empleada en la construcción del modelo.

Gráficamente podemos observar que la calidad de discriminación del modelo es buena puesto que el gráfico de la curva difiere en gran medida de la recta de clasificación aleatoria, validemos esto calculando el área bajo la curva ROC utilizando las siguientes líneas de código en R.

```
# Y: Variable dependiente binaria Activa/No Activa
# Pr: Probabilidad de activación estimada

  library(pROC)
  roc(Y, Pr)
```

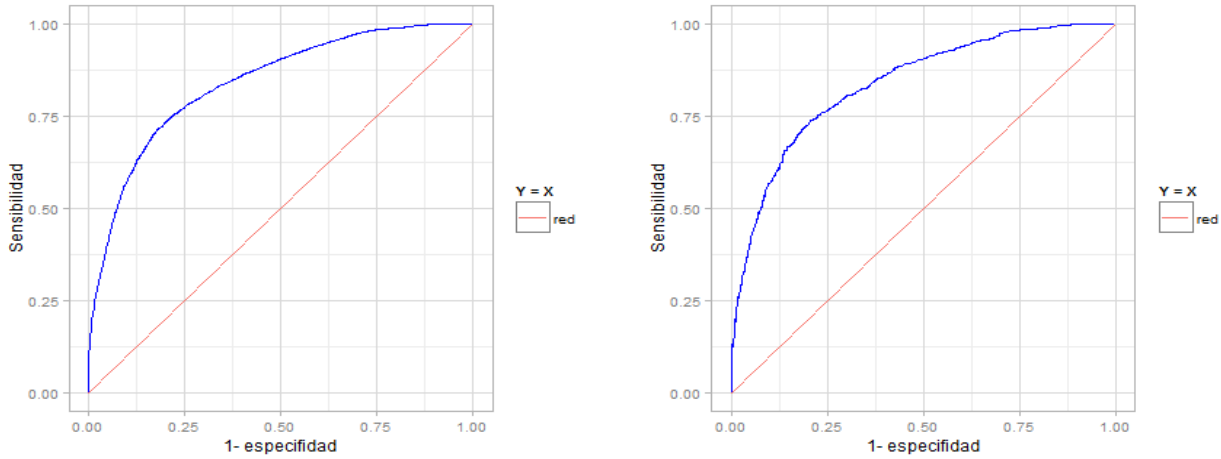

Figura 3.10: Curva ROC muestra de modelamiento-validación.

```
Call:
roc.default(response = Y, predictor = Pr)
Data: Pr in 5031 controls (Y 0) < 10074 cases (Y 1).
área under the curve: 0,84
```

Como resultado tenemos un valor del índice $AUROC$ de $0,84$, lo cual quiere decir que existe una probabilidad de 0,84 de que la probabilidad de activación estimada de un sujeto Activa sea mayor que la probabilidad de activación estimada de un sujeto No Activa elegidos aleatoriamente. Lo anterior nos permite decir que el modelo presenta una alto rendimiento de clasificación cuando el punto de corte varía en el intervalo [0,1].

Adicionalmente presentamos la curva ROC de la probabilidad de activación estimada pero ahora considerando la muestra de validación, para que el modelo se pueda considerar con un alto rendimiento se deben obtener resultados muy similares. En la Figura 3.10b, observamos una curva ROC muy similar a la obtenida utilizando la muestra de modelamiento.

Calculando el área bajo la curva ROC de la Figura 3.10b, se obtuvo un valor de $0,79$, el cual es aproximado al obtenido utilizando la muestra de modelamiento.

- **GINI:** Recordando que el coeficiente de $GINI$ se relaciona con el área bajo la curva ROC mediante la igualdad:

$$GINI = 2 * AUC - 1$$

Se obtuvo un valor para el coeficiente de $GINI$ de $0,68$ en la muestra de modelamiento y un valor de $0,58$ en la muestra de prueba. Considerando los valores para $GINI$ y $AUROC$ podemos concluir que el modelo presenta un excelente rendimiento de clasificación cuando el punto de corte varía en el intervalo [0,1].

Lo anterior se puede validar analizando varias tablas conocidas como tablas de performance, en las cuales se presenta la distribución de los sujetos por los deciles de la probabilidad de activación estimada; en el numeral 3 se analizan a detalle las tablas de performance obtenidas tanto para la muestra de modelamiento y validación.

3. **Tablas de clasificación.** (Tabla de contingencia entre variable dependiente Y y variable pronosticada $\hat{Y}$).

Para generar la tabla de clasificación del modelo se debe primeramente establecer el punto de corte, el cual coincide con el punto donde el KS es máximo o el punto de corte con el cual se obtenga el punto de la curva ROC más cercano al par ordenado (0,1), tal como se mencionó en la descripción de la curva ROC en la sección (2.4.1) del capítulo 2.

Para el presente modelo se obtuvo un punto de corte de 0,698 que corresponde al valor de la probabilidad de activación en el cual se maximiza el $KS = 0,5544$, utilizando este punto de corte y la muestra de modelamiento tenemos como resultado la siguiente tabla de contingencia:

Analizando la columna $\%$ Clasificación correcta de la Tabla 3.12, tenemos que un 71,29 $\%$ y un 82,13 $\%$ de los individuos Activa y No Activa respectivamente son clasificados correctamente. Como podemos ver los porcentajes de

$Y \setminus \hat{Y}$	**Activa**	**No Activa**	% **Clasificación correcta**
Activa	7.182	2.892	71,29 %
No Activa	899	4.132	82,13 %

Tabla 3.12: Tabla de contingencia variable Y y $\hat{Y}$ en muestra de modelamiento.

clasificación correcta son elevados, lo cual nos permite decir que el modelo tiene un elevado poder de clasificación.

Para la muestra de validación utilizando el mismo punto de corte que para la muestra de modelamiento 0,698, se obtiene la Tabla 3.13, de donde tenemos que un 72,30 % y un 80,64 % de los sujetos Activa y No Activa respectivamente fueron clasificados correctamente, estos valores son muy similares a los obtenidos al utilizar la muestra de modelamiento, de esto podemos concluir que el modelo presenta un excelente poder de clasificación.

$Y \setminus \hat{Y}$	**Activa**	**No Activa**	% **Clasificación correcta**
Activa	554	1446	72,30 %
No Activa	779	187	80,64 %

Tabla 3.13: Tabla de contingencia variable Y y $\hat{Y}$ en muestra de validación.

4. **Tablas de performance.**

Las tablas de performance también conocidas como **tablas de desempeño o rendimiento**, son herramientas en las cuales podemos visualizar la calidad de discriminación del modelo por cada decil de la probabilidad de activación estimada. Generalmente para analizar el rendimiento de clasificación del modelo se particiona la probabilidad estimada en 10 intervalos y se analiza el número de sujetos totales, el número de sujetos Activa y No Activa en cada uno de ellos.

En la Tabla 3.14, considerando las notaciones:

- **Min.** Valor mínimo de la probabilidad de activación estimada en cada

intervalo.

- **Max.** Valor máximo de la probabilidad de activación estimada en cada intervalo.
- **Num.** Número de individuos en cada intervalo de la probabilidad de activación estimada.
- %**.** Porcentaje de individuos en cada intervalo de la probabilidad de activación estimada.
- %**Acum.** Porcentaje acumulado de individuos en cada intervalo de la probabilidad de activación estimada.

presentamos los diez intervalos (abierto a la izquierda y cerrado a la derecha) resultantes de la partición de la probabilidad de activación estimada (Prob de activación) en deciles, y la distribución de los sujetos totales, Activa y No Activa en cada uno de ellos.

En lo que corresponde al campo Sujetos totales podemos observar una distribución uniforme de 10 % de los sujetos totales en cada uno de los intervalos de la probabilidad de activación estimada.

Para el campo Sujetos Activa es de esperar que el número de sujetos decrezca estrictamente mientras la probabilidad de activación disminuye, justamente podemos observar este comportamiento analizando la columna Num. Si analizamos la columna $\%Acum$ tenemos que en el rango más alto de la probabilidad estimada se acumula aproximadamente un 15 % de los sujetos Activa, en cambio en el rango más bajo se acumulan únicamente el 1,9 % de los sujetos Activa.

Para el campo Sujetos No Activa en cambio es de esperar que el número de sujetos crezca estrictamente mientras la probabilidad de activación disminuye, precisamente se obtiene este comportamiento si analizamos la columna Num. Si analizamos la columna $\%Acum$ tenemos que en el rango más alto de la probabilidad estimada se acumula únicamente el 0,5 % de los sujetos No Activa,

en cambio en el rango más bajo se acumulan aproximadamente el 26 % de los sujetos No Activa.

Uno de los campos más importantes en una tabla de performance es el correspondiente a la Razón Activa:No Activa, en este campo analizamos el porcentaje de individuos Activa y No Activa respecto al total de individuos en cada uno de los intervalos de probabilidad de activación estimada. El porcentaje de sujetos Activa por cada intervalo (%Act) decrece con la disminución de la probabilidad de activación estimada, en cambio el porcentaje de individuos No Activa por cada intervalo (%No Act) aumenta con la disminución de la probabilidad de activación estimada.

El campo Razón Activa:No Activa se interpreta como sigue, considerando los valores de 98,5 % (Columna %Act) y 1,5 % (Columna %No Act) del intervalo de (0,951,0,999] podemos decir que de todos los sujetos que registran una probabilidad de activación mayor 0,951 y menor o igual que 0,999, el 98,5 % son Activa y el 1,5 % No Activa.

Después de analizar la tabla de performance de la muestra de modelamiento y tras considerar el análisis anterior podemos concluir que el modelo presenta en buen rendimiento de clasificación cuando el punto de corte varía en el intervalo [0,1].

Mediante la utilización de **la muestra de validación** obtenemos la tabla de performance (Tabla 3.15), para que el modelo no esté sobreajustado es de esperar que la distribución de los sujetos totales, Activa, No Activa no varíe significativamente, en nuestro caso en efecto los resultados de la tabla de performance de la muestra de validación son muy similares a los resultados de la tabla de performance de la muestra de modelamiento.

Ahora, pese a que en el desarrollo del modelo descartamos a los sujetos etiquetados como **indeterminados**, esto con el propósito de identificar de mejor

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum	%Act	%No Act
0,951	0,999	1.510	10,0 %	10,0 %	1.487	14,8 %	14,8 %	23	0,5 %	0,5 %	98,5 %	1,5 %
0,920	0,951	1.511	10,0 %	20,0 %	1.412	14,0 %	28,8 %	99	2,0 %	2,4 %	93,4 %	6,6 %
0,890	0,920	1.510	10,0 %	30,0 %	1.364	13,5 %	42,3 %	146	2,9 %	5,3 %	90,3 %	9,7 %
0,846	0,890	1.511	10,0 %	40,0 %	1.324	13,1 %	55,5 %	187	3,7 %	9,0 %	87,6 %	12,4 %
0,758	0,846	1.510	10,0 %	50,0 %	1.195	11,9 %	67,3 %	315	6,3 %	15,3 %	79,1 %	20,9 %
0,609	0,758	1.511	10,0 %	60,0 %	1.020	10,1 %	77,4 %	491	9,8 %	25,1 %	67,5 %	32,5 %
0,514	0,608	1.511	10,0 %	70,0 %	809	8,0 %	85,5 %	702	14,0 %	39,0 %	53,5 %	46,5 %
0,423	0,514	1.510	10,0 %	80,0 %	698	6,9 %	92,4 %	812	16,1 %	55,2 %	46,2 %	53,8 %
0,261	0,423	1.511	10,0 %	90,0 %	578	5,7 %	98,1 %	933	18,5 %	73,7 %	38,3 %	61,7 %
0,001	0,261	1.510	10,0 %	100,0 %	187	1,9 %	100,0 %	1.323	26,3 %	100,0 %	12,4 %	87,6 %
Total		15.105			10.074			5.031				

Tabla 3.14: Tabla de performance muestra de modelamiento (Activa-No Activa).

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum	%Act	%No Act
0,948	0,999	296	10,0 %	10,0 %	291	14,6 %	14,6 %	5	0,5 %	0,5 %	98,3 %	1,7 %
0,919	0,948	297	10,0 %	20,0 %	277	13,9 %	28,4 %	20	2,1 %	2,6 %	93,3 %	6,7 %
0,890	0,919	297	10,0 %	30,0 %	272	13,6 %	42,0 %	25	2,6 %	5,2 %	91,6 %	8,4 %
0,852	0,890	296	10,0 %	40,0 %	259	13,0 %	55,0 %	37	3,8 %	9,0 %	87,5 %	12,5 %
0,784	0,852	297	10,0 %	50,0 %	236	11,8 %	66,8 %	61	6,3 %	15,3 %	79,5 %	20,5 %
0,627	0,783	297	10,0 %	60,0 %	200	10,0 %	76,8 %	97	10,0 %	25,4 %	67,3 %	32,7 %
0,529	0,627	296	10,0 %	70,0 %	169	8,5 %	85,2 %	127	13,1 %	38,5 %	57,1 %	42,9 %
0,446	0,529	297	10,0 %	80,0 %	139	7,0 %	92,2 %	158	16,4 %	54,9 %	46,8 %	53,2 %
0,285	0,446	297	10,0 %	90,0 %	117	5,9 %	98,0 %	180	18,6 %	73,5 %	39,4 %	60,6 %
0,001	0,283	296	10,0 %	100,0 %	40	2,0 %	100,0 %	256	26,5 %	100,0 %	13,5 %	86,5 %
Total		2.966			2.000			966				

Tabla 3.15: Tabla de performance muestra de validación (Activa-No Activa).

manera las características de los sujetos Activa y No Activa, es indispensable realizar las tablas de performance considerando este tipo de sujetos con el fin de ponderar los valores de las columnas %Act y %No Act (campo Razón Activa:No Activa), puesto que pese a que no se consideró a los sujetos indeterminados en el modelamiento estos si son parte de la población total, se espera que la monotonía decreciente y creciente respectivamente se mantenga, en efecto esto se verifica si analizamos la Tabla 3.16, considerando los sujetos indeterminados para la muestra de modelamiento y la Tabla 3.17, considerando los sujetos indeterminados para la muestra de validación.

Después de analizar los resultados obtenidos podemos concluir que el modelo

obtenido presenta un buen ajuste y un alto rendimiento de clasificación y predicción a lo largo del intervalo [0,1].

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa		
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum
0,947	0,999	1.689	10,0 %	10,0 %	1.642	16,3 %	16,3 %	28	0,6 %	0,6 %
0,915	0,947	1.689	10,0 %	20,0 %	1.490	14,8 %	31,1 %	118	2,3 %	2,9 %
0,882	0,915	1.689	10,0 %	30,0 %	1.420	14,1 %	45,2 %	159	3,2 %	6,1 %
0,833	0,882	1.690	10,0 %	40,0 %	1.322	13,1 %	58,3 %	218	4,3 %	10,4 %
0,717	0,833	1.689	10,0 %	50,0 %	1.194	11,9 %	70,2 %	341	6,8 %	17,2 %
0,589	0,716	1.689	10,0 %	60,0 %	915	9,1 %	79,2 %	524	10,4 %	27,6 %
0,502	0,588	1.690	10,0 %	70,0 %	739	7,3 %	86,6 %	675	13,4 %	41,0 %
0,418	0,502	1.689	10,0 %	80,0 %	625	6,2 %	92,8 %	754	15,0 %	56,0 %
0,269	0,418	1.689	10,0 %	90,0 %	532	5,3 %	98,1 %	867	17,2 %	73,2 %
0,001	0,269	1.689	10,0 %	100,0 %	195	1,9 %	100,0 %	1.347	26,8 %	100,0 %
Total		16.892			10.074			5.031		

Prob de activación		Sujetos indeterminados			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	%Act	%No Act
0,947	0,999	19	1,1 %	1,1 %	97,2 %	1,7 %
0,915	0,947	81	4,5 %	5,6 %	88,2 %	7,0 %
0,882	0,915	110	6,2 %	11,8 %	84,1 %	9,4 %
0,833	0,882	150	8,4 %	20,1 %	78,2 %	12,9 %
0,717	0,833	154	8,6 %	28,8 %	70,7 %	20,2 %
0,589	0,716	250	14,0 %	42,8 %	54,2 %	31,0 %
0,502	0,588	276	15,4 %	58,2 %	43,7 %	39,9 %
0,418	0,502	310	17,3 %	75,5 %	37,0 %	44,6 %
0,269	0,418	290	16,2 %	91,8 %	31,5 %	51,3 %
0,001	0,269	147	8,2 %	100,0 %	11,5 %	79,8 %
Total		1.787				

Tabla 3.16: Tabla de performance modelamiento (Activa-No Activa-Indeterminados).

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa		
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum
0,944	0,999	331	10,0 %	10,0 %	319	16,0 %	16,0 %	8	0,8 %	0,8 %
0,915	0,944	332	10,0 %	20,0 %	293	14,7 %	30,6 %	20	2,1 %	2,9 %
0,885	0,915	332	10,0 %	30,0 %	283	14,2 %	44,8 %	29	3,0 %	5,9 %
0,838	0,884	332	10,0 %	40,0 %	257	12,9 %	57,6 %	46	4,8 %	10,7 %
0,747	0,838	331	10,0 %	50,0 %	239	12,0 %	69,6 %	65	6,7 %	17,4 %
0,605	0,746	332	10,0 %	60,0 %	181	9,1 %	78,6 %	103	10,7 %	28,1 %
0,517	0,605	332	10,0 %	70,0 %	152	7,6 %	86,2 %	121	12,5 %	40,6 %
0,437	0,517	332	10,0 %	80,0 %	126	6,3 %	92,5 %	152	15,7 %	56,3 %
0,290	0,437	332	10,0 %	90,0 %	109	5,5 %	98,0 %	158	16,4 %	72,7 %
0,001	0,290	331	10,0 %	100,0 %	41	2,1 %	100,0 %	264	27,3 %	100,0 %
Total		3.317			2.000			966		

Prob de activación		Sujetos indeterminados			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	%Act	%No Act
0,944	0,999	4	1,1 %	1,1 %	96,4 %	2,4 %
0,915	0,944	19	5,4 %	6,6 %	88,3 %	6,0 %
0,885	0,915	20	5,7 %	12,3 %	85,2 %	8,7 %
0,838	0,884	29	8,3 %	20,5 %	77,4 %	13,9 %
0,747	0,838	27	7,7 %	28,2 %	72,2 %	19,6 %
0,605	0,746	48	13,7 %	41,9 %	54,5 %	31,0 %
0,517	0,605	59	16,8 %	58,7 %	45,8 %	36,4 %
0,437	0,517	54	15,4 %	74,1 %	38,0 %	45,8 %
0,290	0,437	65	18,5 %	92,6 %	32,8 %	47,6 %
0,001	0,290	26	7,4 %	100,0 %	12,4 %	79,8 %
Total		351				

Tabla 3.17: Tabla de performance validación (Activa-No Activa-Indeterminados).

Capítulo 4

Automatización de la metodología en R.

4.1. Introducción

Este capítulo lo dedicaremos a describir un algoritmo implementado en el software estadístico R [R Core Team, 2014], el mismo que realiza automáticamente cada uno de los pasos de la metodología explicada en el capítulo 3. Para comprender de mejor manera los códigos utilizados comencemos realizando una introducción a la programación en R.

4.2. R.

R es un lenguaje de programación orientada a objetos libre[1], distribuido bajo la licencia GNU (Acrónimo de General Public License), es utilizado en el manejo y análisis de datos, estadística computacional y generación de gráficos de alta calidad.

Fue inspirado en el software estadístico S[2] desarrollado por John Chambers y colaboradores en Laboratorios Bell (AT&T).

Dentro de la enorme gama de técnicas estadísticas implementados en R podemos enumerar las siguientes, pruebas estadísticas, modelos de ajuste lineales y no

[1]Toda la información acerca del manejo de R, así como su instalador pueden encontrarse en la página web oficial: `http://www.r-project.org`

[2]El nombre S es por la inicial de statistics

lineales, modelos de clasificación, series de tiempo, etc.

Hoy en día es más utilizado que el propio S, debido a su libertad de utilización y distribución, permitiendo a sus usuarios visualizar los códigos en él implementados y modificarlos a conveniencia.

4.3. Descripción de los pasos del algoritmo.

Para la generación del algoritmo a parte del software R-project se utilizó un programa complementario conocido como R AnalyticFlow[3] desarrollado por Ef-prime.

Al igual R, es un software de distribución libre y su principal ventaja es permitir escribir códigos de programación mediante flujogramas (ver Figura 4.1), esto constituye una gran herramienta al momento de desarrollar algoritmos que simulan un determinado proceso.

4.3.1. Paso 1. Filtrado de variables explicativas.

En este paso el algoritmo, para cada una de las variables explicativas continuas disponibles calcula el estadístico de Kolmogorov-Smirnov, Anderson Darling y el coeficiente de correlación de Pearson (explicados a detalle en la sección (2.1) del capítulo 2) mediante la utilización de la siguiente función programada en R.

```
# FUNCION:ind_var_num
# Y: Variable dependiente, variable: variable explicativa
ind_var_num <-function (Y,variable){
  require(dgof)
  require(kSamples)
  # Activa=1 - No Activa=0
    vars <- data.frame(Y,variable)
    vars_A <- subset(vars,subset=vars[,1]==1)
```

[3]R AnalyticFlow se puede descargar de su página oficial: http://www.ef-prime.com/products/ranalyticflow_en

```
    vars_NA <- subset(vars,subset=vars[,1]==0)
  # Estadistico KS
    ks <- dgof::ks.test(vars_A[,2],vars_NA[,2],alternative="two.sided")
    ks <- round(as.numeric(ks$statistic),4)
  # Estadistico AD
    if(ks>0){
        AD <- ad.test(vars_NA[,2],vars_A[,2],method = c("asymptotic"))[7]
        AD <- as.data.frame(AD)[1,1]
    }else{
      AD <- 0}
  # Coef de correlación de pearson
    corr <- round(abs(as.numeric(cor.test(vars[,1],vars[,2],
                    method="pearson",conf.level = 0.95,
                    na.action=getOption("na.action"))[4])),4)
  return(c(ks,AD,corr))
}
```

En lo que se refiere a las variables explicativas categóricas disponibles, el algoritmo calcula el coeficiente de contingencia de Pearson, el valor de información, la diferencia de información por categoría (Explicados a detalle en la sección (2.2) del capítulo 2), básicamente las funciones en R que permiten realizar lo descrito anteriormente se presentan a continuación.

```
# Raiz de la media de los residuos al cuadrado
  RMR <- function(x,y,pesos=NULL){
    if(!is.null(pesos)){
      rmr <- sqrt(sum((pesos/sum(pesos))*((x-y)^2)))
    }else{
      rmr <- sqrt(sum(((x-y)^2))/length(x))
    }
    return(rmr)
  }
# Valor de información (IV)  x:Activa=1 - y:No Activa=0
  VI <- function(x,y){
```

```
      aux <- ifelse(x/sum(x)==0,1,x/sum(x))
      wof <- log((y/sum(y))/aux)
      wof <- ifelse(wof==-Inf,0,wof)
      VI <-   sum(((y/sum(y))-(x/sum(x)))*wof)
    return(VI)
  }
#  FUNCION: ind_var_cat
ind_var_cat <- function (Y,variable){
      vars <- data.frame(Y,variable)
    # Prueba de independencia - Coef de contingencia
      chisq<-chisq.test(Y,variable)
      chisq<- as.numeric(chisq$statistic)
      cont <- round(sqrt(chisq/(chisq+length(variable))),4)
    # DIC
       tc <- table(Y,variable)
       tot <- ifelse ((tc[1,] + tc[2,])==0,1,(tc[1,] + tc[2,]))
       odds <- tc[2,]/tot
       p <- as.vector (table(Y))
       porc_m <- rep(p[2]/sum(p),dim(tc)[2])
       dic <- round(RMR(odds,porc_m,pesos=tot),4)
    # Valor de información
       vi <- VI(tc[1,],tc[2,])
  return(c(cont,dic,vi))
}
```

Las funciones detalladas anteriormente son implementadas en R AnalyticFlow, tal como muestra la Figura 4.1. En este flujograma realizamos las siguientes tareas:

1. CARGAR BDD. Empezamos cargando la base de datos, la cual debe contener la variable dependiente Y (Activa/No Activa) y todas las variables explicativas disponibles ya sean numéricas o categóricas.

2. POBLACIÓN. En ocasiones es necesario realizar un filtrado previo de registros, tal como lo hicimos en el capítulo 3, cuando excluimos a los sujetos no

BANCARIZADOS, en este punto se obtiene la población que se utilizará en el modelamiento.

3. VAR PREDICTIVAS. Se cargan al área de trabajo todas las funciones necesarias para el cálculo de las medidas KS, AD, VI, etc.
4. FILTRO. Se ejecutan las funciones cargadas en 3 a la base de datos disponible.
5. NUMÉRICAS. Se muestran los resultados obtenidos para las variables numéricas.
6. CATEGÓRICAS. Se muestran los resultados obtenidos para las variables categóricas.
7. SIN INFO. Se muestran aquellas variables que presentan un porcentaje de valores perdidos superior al $50\,\%$.

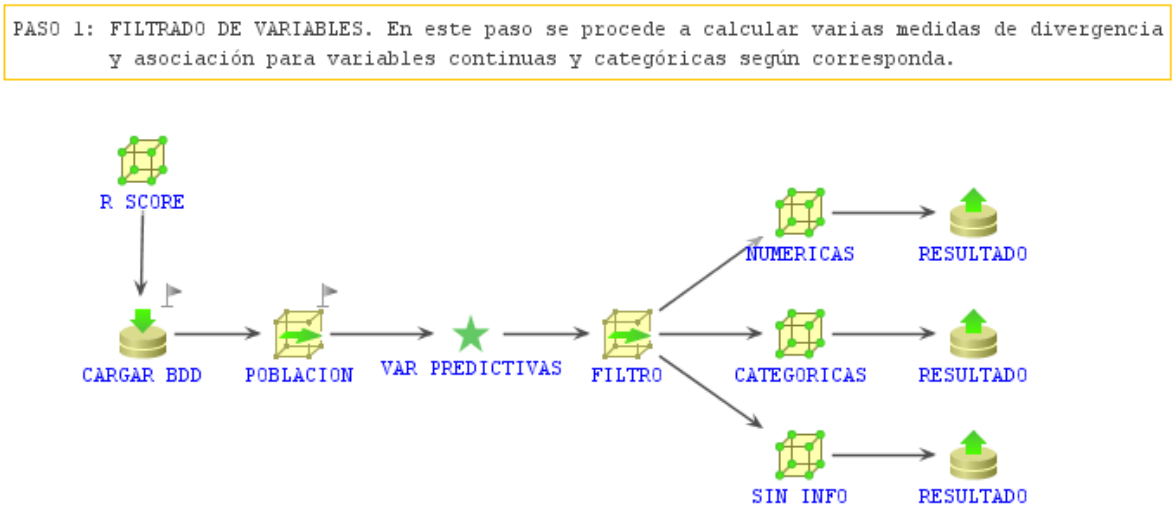

Figura 4.1: Flujograma del algoritmo correspondiente al paso 1.

4.3.2. Paso 2. Selección de variables explicativas continuas.

En este paso se procede a seleccionar que variables ingresarán al modelo de regresión logística como variables continuas, es decir aquellas que no se discretizarán mediante árboles de decisión.

Varios son los criterios considerados para la selección de las variables continuas, en este estudio consideramos los siguientes:

1. **Poder predictivo:** En lo que se refiere al poder predictivo utilizaremos las medidas de divergencia calculadas en el paso 1, entre las cuales tenemos KS, AD, $|\text{CORR}|$, para seleccionar las variables numéricas continuas con los valores en las medidas de divergencia más elevados, estas medidas se calculan en el paso 1 del algoritmo.

2. **Información:** Al momento de ingresar una variable numérica continua a un modelo de regresión se debe plantear la inquietud acerca de las observaciones que presenta esta variable, pues por ejemplo una variable con el $80\,\%$ de sus observaciones iguales a una constante (por ejemplo 0) no contribuiría en la predicción de la probabilidad.

Un posible código generado en R que realice la tarea correspondiente al análisis de la información de la variable es el siguiente:

```
# index_var: Vector con el número de columna de las variables
# continuas en la base de datos BDD
# n_cont: Número de variables continuas a seleccionar
  cont_var <- c(0)
  for(k in 1:length(index_var)){
   cuts <- quantile(BDD[,index_var[k]],seq(0.1,0.9,0.1),na.rm=TRUE)
   n_cut <- length(cuts)
   if(!all(cuts[1:(n_cut-2)]==cuts[2:(n_cut-1)])){
      cont_var <- c(cont_var,index_var[k])
     }else{
      cont_var <- c(cont_var,-999)
    }
  }
  cont_var <- cont_var[cont_var > 0]
  var_continuas <- BDD[,cont_var][,1:n_cont]
```

En el flujograma de la Figura 4.2, se implementó la tareas correspondientes a la información de las variables continuas, pues lo correspondiente al poder predictivo se realiza ya en el paso 1. El paso 2 está compuesto por las siguientes tres acciones:

1. BDD. Se carga la base con las variables numéricas disponibles.

2. VAR CONTINUAS. Se ejecuta el código encargado de descartar aquellas variables continuas con un porcentaje de observaciones iguales a una constante superior al 80 %.

3. RESULTADO. Se presenta la base de variables numéricas continuas seleccionadas.

```
PASO 2: SELECCIÓN DE VARIABLES CONTINUAS. En este paso se procede a obtener las variables
        continuas candidatas a ingresar al modelo logístico, es decir estas variables no
        se discretizarán mediante árboles de decisión.
```

Figura 4.2: Flujograma del algoritmo correspondiente al paso 2.

4.3.3. Paso 3. Generación de dummys y probabilidades de Activa.

Este es uno de los pasos más importantes del algoritmo y el más complicado de implementar, en este paso se realiza la interacción entre la variable dependiente y una variable explicativa, con el propósito de construir las variables dummy y probabilidad de activa asociadas a dicha interacción; mediante el código en R que a continuación se describe.

```
# Y: variable dependiente binaria
# BDD: Base de datos de variables filtradas en el paso 1 y 2
# x1: i-ésima variable explicativa en base BDD
  require(party)
  Ya <- ifelse(Y==1,"MALO","BUENO")
  d <- data.frame(Ya,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  assign("x1",d[,i+1],envir=.GlobalEnv)
  arb<-ctree(y~x1,data=d)
  plot(arb,main =paste("x1 =",names(BDD)[i]))
```

El algoritmo adicionalmente realiza la interacción entre la variable dependiente y dos variables explicativas, el código en R que se podría utilizar para realizar esta tarea es el siguiente.

```
# Y: variable dependiente binaria
# BDD: Base de datos de variables filtradas en el paso 1 y 2
# x1: i-ésima variable explicativa en base BDD
# x2: j-ésima variable explicativa en base BDD
  require(party)
  Ya <- ifelse(Y==1,"MALO","BUENO")
  d <- data.frame(Ya,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  assign("x1",d[,i+1],envir=.GlobalEnv)
  assign("x2",d[,j+1],envir=.GlobalEnv)
  arb<-ctree(y~x1+x2,data=d)
  plot(arb,main =paste("x1 =",names(BDD)[i],"y" ,"x2 =",names(BDD)[j]))
```

Para el caso de las variables categóricas se realiza la interacción entre la variable dependiente y una variable explicativa, entre la variable dependiente y dos variables explicativas, entre la variable dependiente y tres variables explicativas.

Antes de proceder a ejecutar las tareas del flujograma de la Figura 4.3, las cuales en resumen consisten en graficar los árboles de decisión y calcular tanto las variables dummys como las probabilidades de éxito, el algoritmo permite ingresar por el teclado el número de variables que deseamos considerar, en base al gráfico de sedimentación de la Figura 4.3, el mismo que es construido con las medidas calculadas en el paso 1 (La manera de obtener el gráfico de sedimentación de la Figura 4.3, se explicó a detalle en la sección (3.1.5) del capítulo 3).
Las tareas realizadas mediante el flujograma de la Figura 4.4, son:

1. 1 VARIABLE. Se cargan al área de trabajo las funciones necesarias para la realización de los gráficos de los árboles y el cálculo automático de las variables dummys y probabilidades de éxito para la interacción de la variable dependiente y una variable explicativa numérica.

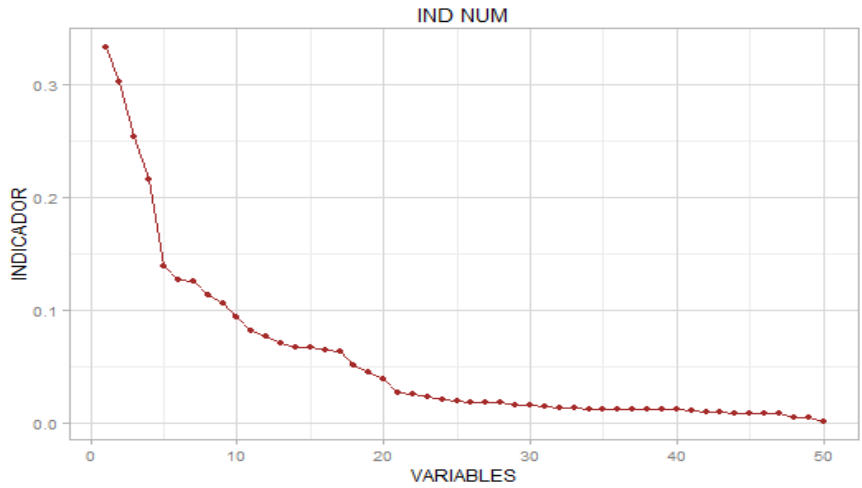

Figura 4.3: Flujograma del algoritmo correspondiente al paso 2.

- GRAFICAR. Se crea un archivo en formato pdf con el gráfico de los árboles de una variable.
- DUMMYS PRBE. Se calculan las variables dummys y las probabilidades de éxito para los árboles de una variable.

2. 2 VARIABLES. Se cargan al área de trabajo las funciones necesarias para la realización de los gráficos de los árboles y el cálculo automático de las variables dummys y probabilidades de éxito para la interacción entre la variable dependiente y dos variables explicativas numéricas.

   - GRAFICAR. Se crea un archivo en formato pdf con el gráfico de los árboles de dos variables.
   - DUMMYS PRBE. Se calculan las variables dummys y las probabilidades de éxito.

3. 1, 2, 3 VARIABLES CAT. Se cargan al área de trabajo las funciones necesarias para la realización de los gráficos de los árboles y el cálculo automático de las variables dummys y probabilidades de éxito para la interacción entre la variable dependiente y una, dos, tres variables explicativas categóricas.

   - GRAFICAR. Se crea 3 archivos en formato pdf cada uno con los gráficos de los árboles de 1, 2, 3 variables categóricas respectivamente.

- DUMMYS PRBE. Se calculan las variables dummys y las probabilidades de éxito para los árboles de 1, 2, 3 variables categóricas.

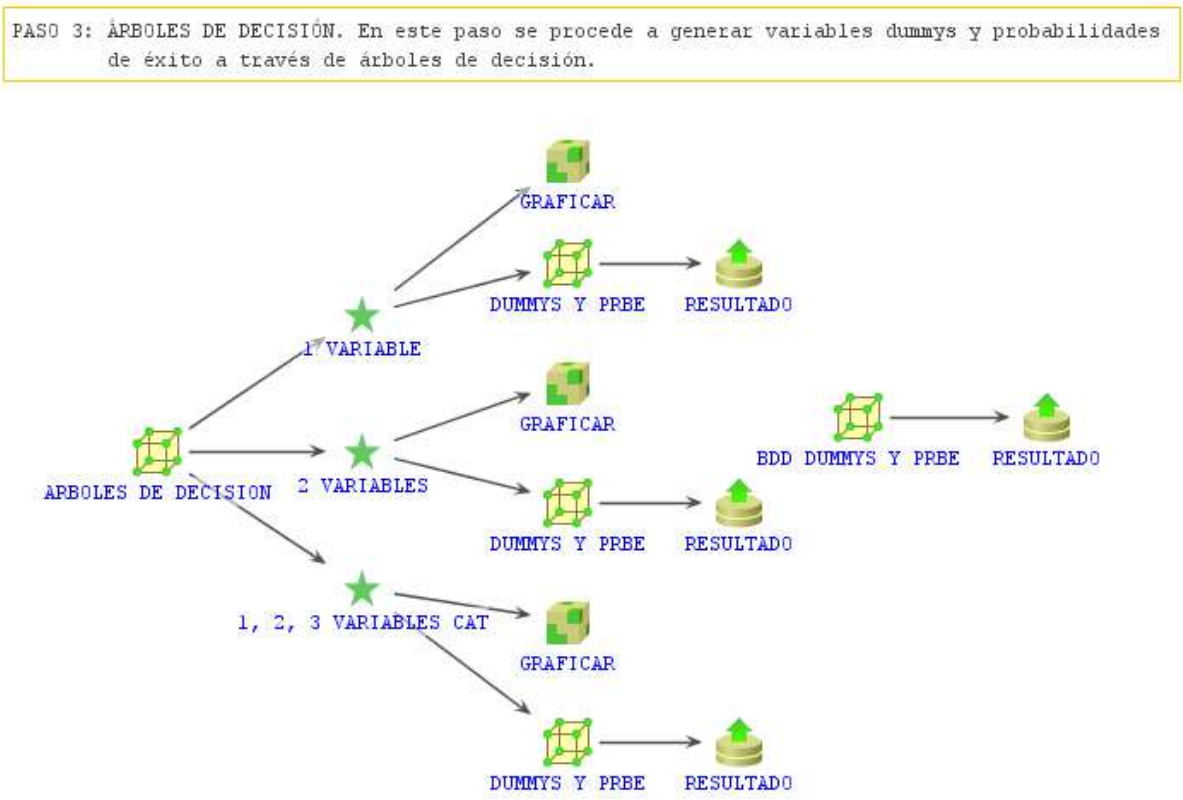

Figura 4.4: Flujograma del algoritmo correspondiente al paso 3.

4.3.4. Paso 4. Filtrado de variables dummys y probabilidades de Activa.

Este paso del algoritmo es similar al paso 1, solo que este caso calculamos las medidas KS, coeficiente de correlación de Pearson, Coeficiente de contingencia de Pearson, valor de información para las variables construidas en el paso 3 mediante los árboles de decisión.

Adicionalmente se realiza un filtrado extra, el cual consiste en tomar en parejas las variables explicativas con una correlación superior a 0,7, y seleccionar únicamente aquella con medidas KS, $|\text{CORR}|$, $CONT$, VI más elevadas, un posible código que realice esta tarea es el siguiente.

```
# BDD_D_PRBE: Base de variables dummys y probabilidades de éxito
#             ordenadas decrecientemente a través de las medidas
#             de divergencia.
# BDD_D_PRBE_FINAL: Base de variables dummys y probabilidades de
#                   éxito final.
  AUX <- cor(BDD_D_PRBE)
  pos <- which(((abs(AUX)>=0.7) & (row(AUX) < col(AUX))),arr.ind=T)
  col_elim <- numeric(nrow(pos))
  for(i in seq(1:nrow(pos))){
      aux_col_elim <- c(pos[i,1],pos[i,2])
      if (!any(col_elim %in% aux_col_elim)){
         col_elim [i] <- pos[i,which.max(c(pos[i,1],pos[i,2]))]
        }
   }
 col_elim <- unique(col_elim[col_elim>0])
BDD_D_PRBE_FINAL <- BDD_D_PRBE[,-(col_elim)]
```

A continuación se describen las tareas realizadas en este paso, las cuales son implementadas en el flujograma de la Figura 4.5.

1. BDD DUMMYS-PRBE. Empezamos cargando la base de datos de las variables dummys y probabilidades de éxito generadas en el paso 3.

2. VAR PREDICTIVAS. Se cargan al área de trabajo todas las funciones indicadas en el paso 1, para calcular las medidas de divergencia y asociación de las variables dummys y probabilidades de éxito.

3. FILTRO 1. Se ejecutan las funciones cargadas en el numeral 2 para la base de variables generada en el paso 3.

4. DUMMYS Y PRBE. Se emparejan las variables explicativas, se calcula su correlación, y para aquellas con una correlación superior a 0,7 se selecciona aquella con las medidas divergencia y asociación más altas.

5. FILTRO 2. Se ejecutan las funciones cargadas en el numeral 2, pero en este caso para la base resultante de la eliminación de las variables altamente

correlacionadas.

6. DUMMYS PRBE FINAL. Nuevamente a través de un gráfico de sedimentación, pero esta vez para las variables dummys y probabilidad de éxito, seleccionamos las variables que utilizaremos para el ajuste del modelo de regresión.

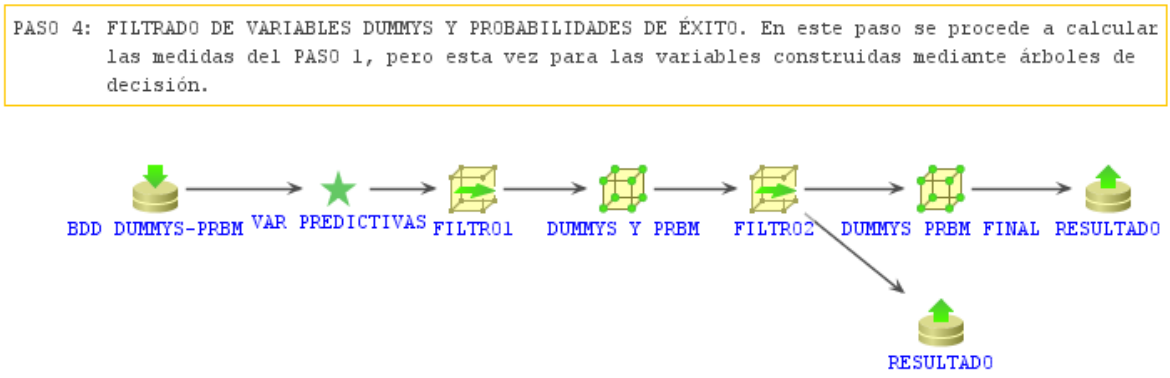

Figura 4.5: Flujograma del algoritmo correspondiente al paso 4.

4.3.5. Paso 5. Ajuste del modelo de regresión logística y validación.

En este paso se procede a ajustar un modelo de regresión logística a las variables generadas en el paso 2 y en el paso 4. Para obtener los regresores que conformen el modelo final se utiliza el método de pasos hacia atrás en base al criterio de Akaike, es decir empezamos considerando todas las variables y procedemos a probar la eliminación de los regresores de tal manera que se minimice el valor de AIC.

Un posible código en R para realizar lo descrito anteriormente estaría dado por las siguientes líneas.

```
# Y: variable dependiente binaria
# base_regresion: Base de variables para la regresión con la variable
# dependiente Y en la primera columna.
# Y debe ser tipo factor para la regresión
  Y <- base_regresion[,1]
  Y_f <-as.factor(Y)
  base_regresion <- data.frame (Y_f,base_regresion[,-1])
```

```
# Partimos de la regresión inicial con todas las variables:
  ro <- glm (Y_f~., data=base_regresion[,-1], family = binomial("logit"))
# Seleccionamos mejor modelo en base al AIC:
  mejor_r <<- stepAIC (ro, direction = "backward")
```

Una vez obtenido el modelo de regresión logística se procede a presentar varios resultados importantes en la validación del mismo. Se presentan resultados tales como el coeficiente de determinación R^2, las variables que conforman el modelo final con su respectivo coeficiente, error estándar, intervalo de confianza, significancia, etc.

A continuación describimos las tareas que realiza el flujograma de la Figura 4.6.

1. BDD REGRESIÓN. Empezamos cargando la base de datos de las variables continuas, dummys y probabilidades de éxito generadas en los pasos 2 y 4.
2. LIBRERÍAS. Se cargan al área de trabajo todas las librerías (Paquetes de R) necesarias para ejecutar la función de regresión logística.
3. FUNCIÓN REG LOGÍSTICA. Se cargan al área de trabajo el código de las funciones necesarias para ajustar el modelo de regresión y la obtención de los resultados mencionados anteriormente.
4. REG LOGÍSTICA. Se ejecutan las funciones cargadas en el numeral 3.
5. RESULTADO REGRESIÓN. En este punto se presentan los resultados obtenidos, entre los más importantes podemos enumerar los siguientes:
   - R^2
   - Variables en la regresión logística.
   - Matriz de correlación.
   - Índice de condicionamiento (Problema de multicolinealidad).
6. RESULTADO SCORE (Probabilidad de activación). Una vez validado estadísticamente el modelo de regresión se procede a evaluar la calidad de predicción y discriminación, en este punto nos centramos en analizar la probabilidad de éxito pronosticada (Activación en el presente trabajo),

básicamente se calcula el estadístico de Kolmogorov-Smirnov, el área bajo la curva ROC, el coeficiente de $GINI$, tabla de contingencia entre la variable dependiente real y la pronosticada, tabla de performance, y algunos gráficos en los cuales se puede visualizar la calidad de discriminación entre los casos éxito y fracaso (Activa / No Activa).

7. FUNCIÓN RESULTADOS. Se ejecuta la función para la obtención de los resultados mencionados en el numeral 6.

8. RESULTADOS. Se presentan los resultados mencionados en el numeral 6.

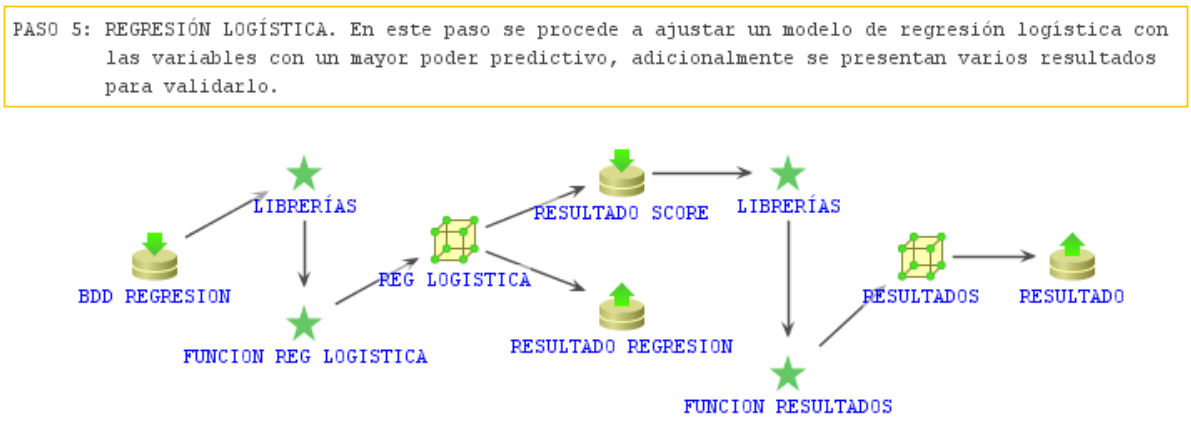

Figura 4.6: Flujograma del algoritmo correspondiente al paso 5.

4.4. Flujograma del algoritmo implementado en R.

En esta sección presentamos el flujograma total del algoritmo implementado en R, es decir presentamos en conjunto los pasos descritos en la sección (4.3). Cabe recalcar que el algoritmo puede ser utilizado por usuarios que no tengan conocimiento de R, puesto que únicamente realizando click sobre los nodos de la Figura 4.7, las líneas de código se van ejecutando, sin la necesidad de visualizarlas.

Las líneas de código del algoritmo descrito en la sección (4.3) se presenta en su totalidad y a detalle en el Anexo D.

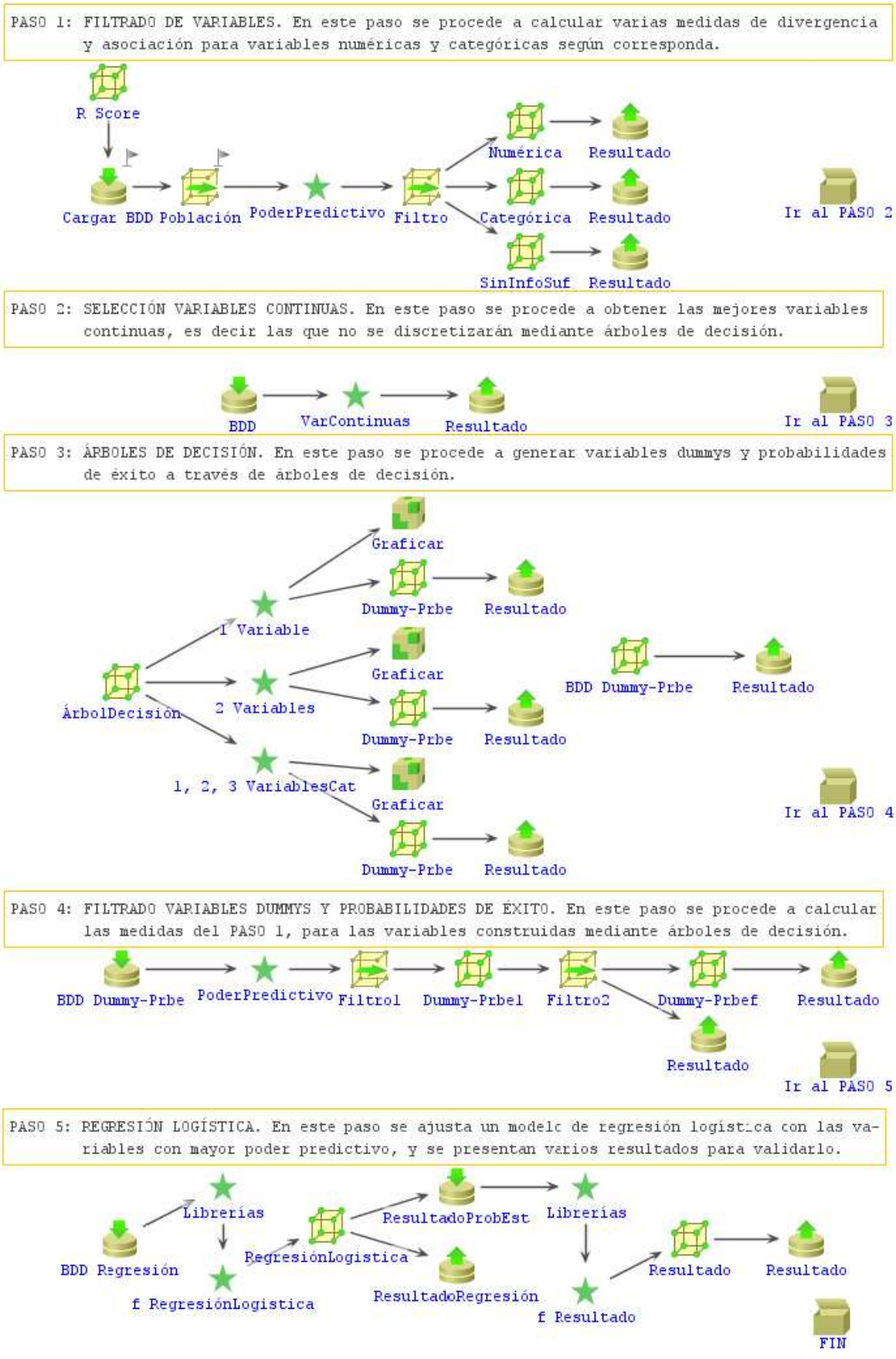

Figura 4.7: Flujograma del algoritmo implementado en R.

4.5. Resultados obtenidos a partir de la ejecución del algoritmo en R.

4.5.1. Modelo de regresión logística.

En esta sección presentamos el modelo de regresión logística final, obtenido tras la ejecución del algoritmo en R. En la Tabla 4.1, se presenta el nombre del regresor, el parámetro estimado (Coeficiente), error estándar, significancia, los extremos del intervalo de confianza (Lim inf y Lim sup) que conforman el modelo obtenido, el mismo que presenta un coeficiente de determinación $R^2 = 0,4779$, análogamente que para el modelo manual, podemos decir que considerando el tipo de información disponible, el modelo presenta un ajuste aceptable.

i	Variable	Coeficiente	Error estándar	Significancia	Lim inf	Lim sup
1	*Porc_Cupo_Util*	0,010	0,001	0,000	0,007	0,013
2	*Num_TC_Consumo*	0,172	0,028	0,000	0,116	0,228
3	*Max_Antiguedad_TC*	-0,001	0,001	0,009	-0,002	0,000
4	*Mejor_TC*	-0,329	0,038	0,000	-0,406	-0,253
5	*prba_TC_Abiert_Ult_3M y Copen_Vig_3_all*	2,519	0,156	0,000	2,212	2,826
6	*prba_Ln_Cupo_Prom_TC_Vig y Copen_Vig_3_all*	0,871	0,180	0,000	0,517	1,226
7	*prba_Ln_Max_Cupo_Vig_TC y TC_Abiert_Ult_3M*	3,285	0,173	0,000	2,948	3,629
8	*d_Porc_Cupo_Util_6M y Ln_Cupo_Prom_TC_Vig*	0,831	0,073	0,000	0,687	0,974
9	*d_Cupo_Prom_TC_Consumo y Ln_Cupo_Prom_TC_Vig*	0,540	0,049	0,000	0,442	0,637
10	*d_Max_Antiguedad_TC y Ln_Cupo_Prom_TC_Vig*	0,680	0,065	0,000	0,553	0,809
11	*Constante*	-2,189	0,348	0,000	-2,873	-1,507

Tabla 4.1: Variables explicativas en el modelo automático.

De la Tabla (4.1), podemos concluir que las variables consideradas son significativas y considerando el significado de cada variable (Ver ANEXO A) tenemos que los signos de los coeficientes son correctos, pues las variables:

1. *Porc_Cupo_Util*. Porcentaje de cupo utilizado.

2. *Num_TC_Consumo*. Número de tarjetas con consumo.

3. *prba_TC_Abiert_Ult_3M_y_Copen_Vig_3M*. Probabilidad de activa de la interacción entre la variable número de tarjetas de crédito abiertas los últimos 3 meses anteriores al punto de observación y la variable cantidad de operaciones abiertas que se encuentren vigentes (que no hayan sido canceladas) en los últimos 3 meses anteriores al punto de observación.

4. *prba_Ln_Cupo_Prom_TC_Vig_y_Copen_Vig_3M*. Probabilidad de activa de la interacción entre la variable logaritmo natural del cupo promedio de las tarjetas vigentes al punto de observación y la variable cantidad de operaciones abiertas que se encuentren vigentes en los últimos 3 meses anteriores al punto de observación.

5. *prba_Ln_Max_Cupo_Vig_TC_y_TC_Abiert_Ult_3M*. Probabilidad de activa de la interacción entre la variable logaritmo natural del cupo máximo de las tarjetas vigentes al punto de observación y la variable número de tarjetas de crédito abiertas los últimos 3 meses anteriores al punto de observación.

6. *d_Porc_Cupo_Util_6M_y_Ln_Cupo_Prom_TC_Vig*. Variable dummy construida mediante la interacción entre la variable porcentaje de cupo utilizado durante los últimos 6 meses anteriores al punto de observación y la variable logaritmo natural del cupo promedio de las tarjetas vigentes al punto de observación.

7. *d_Cupo_Prom_TC_Consumo_Ln_Cupo_Prom_TC_Vig*. Variable dummy construida mediante la interacción entre la variable cupo promedio de las tarjetas que registren consumos al punto de observación y la variable logaritmo natural del cupo promedio de las tarjetas vigentes al punto de observación.

8. *d_Max_Antiguedad_TC_y_Ln_Cupo_Prom_TC_Vig*. Variable dummy construida mediante la interacción entre la variable tiempo transcurrido desde que el sujeto aperturó la primera tarjeta y la variable logaritmo natural del cupo promedio de las tarjetas vigentes al punto de observación.

Deben ingresar al modelo a premiar (A valores más altos de la variable mayor probabilidad de activación), es decir su respectivo coeficiente debe ser positivo, y las variables:

1. $Max\_Antiguedad\_TC$. Tiempo transcurrido desde que el sujeto aperturó la primera tarjeta de crédito
2. $Mejor\_TC$. Variable numérica que cuantifica el tipo de tarjeta de crédito.

En cambio deben ingresar a castigar (A valores más altos de la variable menor probabilidad de activación), es decir su respectivo coeficiente debe ser negativo.

La forma de cálculo de todas las variables mencionadas anteriormente se explicó a detalle en la sección (3.1.6) del capítulo 3.

En las secciones siguientes analizaremos la calidad de discriminación y predicción del modelo de regresión logística obtenido mediante la ejecución del algoritmo.

4.5.2. Medidas de calidad de discriminación.

En la sección validaremos la calidad de discriminación del modelo calculando los indicadores de calidad de discriminación y predicción siguientes:

1. KS.
2. Área bajo la curva ROC.
3. $GINI$.

Considerando la Tabla 4.2, tenemos que el modelo estimado por el algoritmo presenta un KS de 0,5726, una área bajo la curva ROC de 0,8635 (Ver Figura 4.8), y un coeficiente de $GINI$ de 0,7270, con lo que podemos concluir que posee una buena calidad de discriminación y predicción, es decir el modelo se ajusta en gran medida a los datos de modelamiento.

i	Indicador	Valor
1	KS	0,5726
2	AUC	0,8635
3	$GINI$	0,7270

Tabla 4.2: Indicadores del modelo obtenido de la ejecución del algoritmo en R.

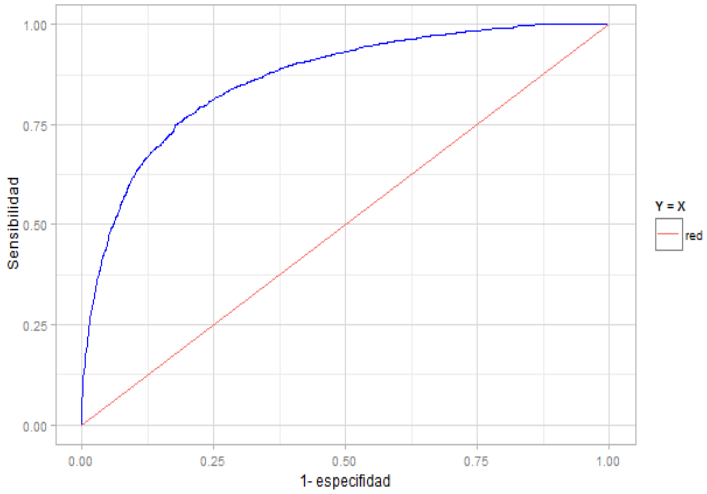

Figura 4.8: Curva ROC del modelo obtenido de la ejecución del algoritmo en R.

4.5.3. Tabla de contingencia entre la variable Y y $\hat{Y}$.

En la Tabla 4.3, se presenta la tabla de contingencia entre la variable dependiente Y y la variable dependiente pronosticada $\hat{Y}$.

Utilizando el punto de corte igual a 0,601 que corresponde al valor de la probabilidad de activación en el cual se maximiza el $KS = 0,5726$, tenemos que un 89,0 % y un 62,0 % de los sujetos Activa y No Activa respectivamente fueron clasificados correctamente.

$Y \setminus \hat{Y}$	**Activa**	**No Activa**	**% Clasificación correcta**
Activa	8.973	1.101	89,0 %
No Activa	1.907	3.124	62,0 %

Tabla 4.3: Tabla de clasificación del modelo obtenido a través del algoritmo en R.

4.5.4. Tablas de performance.

Presentamos ahora la tabla de performance del modelo (Tabla 4.4), en la cual observamos la distribución de los sujetos totales, sujetos Activa, sujetos No Activa, por cada uno de los diez intervalos de la probabilidad de activación.

Podemos observar que el porcentaje de sujetos Activa en cada rango de la probabilidad de activación estimada disminuye con la disminución de la probabilidad, adicionalmente presenta un decrecimiento estricto. El porcentaje de sujetos No Activa en cada rango de la probabilidad de activación estimada en cambio aumenta con la disminución de la probabilidad estimada y presenta un crecimiento estricto.

Algo adicional que podemos mencionar es que en el rango más alto de la probabilidad estimada el 98,2 % de los sujetos son Activa y únicamente el 1,8 % son No Activa, en cambio en el rango más bajo de la probabilidad estimada únicamente el 10 % de los sujetos son Activa y el 88,2 % son No Activa.

De lo anterior podemos concluir que el modelo presenta un alto rendimiento de discriminación a lo largo del intervalo [0,1].

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum	%Act	%No Act
0,805	0,999	1.531	10,0 %	10,0 %	1.503	14,9 %	14,9 %	28	0,6 %	0,6 %	98,2 %	1,8 %
0,619	0,805	1.493	10,0 %	20,0 %	1.425	14,1 %	29,1 %	68	1,4 %	1,9 %	95,4 %	4,6 %
0,477	0,619	1.527	10,0 %	30,0 %	1.412	14,0 %	43,1 %	115	2,3 %	4,2 %	92,5 %	7,5 %
0,360	0,477	1.505	10,0 %	40,0 %	1.325	13,2 %	56,2 %	180	3,6 %	7,8 %	88,0 %	12,0 %
0,255	0,360	1.507	10,0 %	50,0 %	1.224	12,2 %	68,4 %	283	5,6 %	13,4 %	81,2 %	18,8 %
0,157	0,255	1.503	10,0 %	60,0 %	1.068	10,6 %	79,0 %	435	8,6 %	22,0 %	71,1 %	28,9 %
0,097	0,157	1.520	10,0 %	70,0 %	862	8,6 %	87,5 %	658	13,1 %	35,1 %	56,7 %	43,3 %
0,054	0,097	1.505	10,0 %	80,0 %	645	6,4 %	93,9 %	860	17,1 %	52,2 %	42,9 %	57,1 %
0,026	0,054	1.506	10,0 %	90,0 %	432	4,3 %	98,2 %	1.074	21,3 %	73,6 %	28,7 %	71,3 %
0,010	0,026	1.508	10,0 %	100,0 %	178	1,8 %	100,0 %	1.330	26,4 %	100,0 %	11,8 %	88,2 %
Total		15.105			10.074			5.031				

Tabla 4.4: Tabla de performance del modelo obtenido a través del algoritmo en R.

Después de analizar los resultados obtenidos tras la ejecución del algoritmo podemos concluir que el modelo obtenido presenta un buen ajuste y un alto rendimiento de

clasificación y predicción a lo largo del intervalo [0,1].

En el capítulo siguiente nos centraremos en analizar las posibles ventajas y desventajas del algoritmo implementado en R-project, es decir compararemos los resultados obtenidos tras la realización manual de la metodología y la ejecución del algoritmo y el tiempo utilizado en la realización de cada una de ellas.

4.5.5. Multicolinealidad.

Para estudiar el problema de multicolinealidad en el modelo de regresión obtenido tras la ejecución del algoritmo en R, de forma análoga que en el modelo manual, utilizaremos el índice de condicionamiento (IC), en este caso tenemos:

$$IC = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{2,25}{0,21}} = 3,25$$

para el modelo automático se tiene un $IC = 3,25 < 10$, lo que implica que no hay presencia de multicolinealidad.

Capítulo 5

Comparación de resultados obtenidos.

La metodología estudiada en el capítulo 3 es ampliamente utilizada en la construcción de modelos de predicción de una variable dependiente binaria, de ahí surgió la necesidad de implementar cada uno de sus pasos en un lenguaje de programación y generar un algoritmo capaz de recibir las variables explicativas y obtener la predicción correspondiente.

El propósito principal de la implementación fue disminuir el tiempo de trabajo, puesto que en la metodología existen varias tareas tediosas y repetitivas que consumen en gran cantidad el recurso tiempo y que con el conocimiento de programación adecuado resulta sencillo implementarlas.

Se debe considerar que el algoritmo generado no puede competir con los factores conocimiento y experiencia del negocio, es por esta razón que el algoritmo se debe utilizar únicamente para tener una idea de las variables que se deben incluir y cuales de ellas son las que mejor predicen la variable dependiente binaria, es decir, el ser humano es quien al final tomará las decisiones.

En el presente capítulo estudiamos las ventajas y desventajas que ocasiona el generar el modelo de activación de tarjetas de crédito con el algoritmo implementado en R, que en adelante lo denominaremos modelo automático, comparándolo con el modelo resultante de la realización manual de cada uno de los pasos de la metodología, que en adelante se denominará método manual.

5.1. Tiempo de ejecución.

Cuando se desarrollan tares tediosas y repetitivas se consume una gran cantidad de tiempo, el objetivo de la implementación de estas tareas es disminuir el tiempo que tardaríamos en realizarlas manualmente. El presente estudio tiene como objetivo justamente implementar la metodología utilizada en la construcción de modelos de predicción cuando la variable dependiente es binaria, y analizar su funcionamiento mediante la generación de un modelo de activación de tarjetas de crédito.

En la construcción de las líneas de código del algoritmo se utilizó un tiempo de 30 días dedicados a su diseño y de 15 días utilizados en la realización de pruebas y ajustes adicionales. El tiempo de 45 días de generación del código queda plenamente justificado si analizamos una de las principales ventajas de la implementación, la cual consiste en la disminución considerable del tiempo en la construcción del modelo.

En la Tabla 5.1, presentamos los tiempos utilizados en la construcción del modelo mediante la ejecución del algoritmo implementado en R (modelo automático) y mediante la realización manual de cada uno de los pasos de la metodología (modelo manual).

Cabe recalcar que el tiempo de ejecución del algoritmo puede ser disminuido dependiendo del procesador de la máquina en la cual se compile. El tiempo de la Tabla 5.1, es el obtenido al utilizar un procesador común y corriente (procesador I_5).

Modelo	**Tiempo**
Modelo automático	4 horas
Modelo manual	40 horas

Tabla 5.1: Tiempo utilizado en la generación de los modelos.

Analizando los datos sobre el tiempo utilizado en la generación de los modelos (Tabla 5.1) tenemos que al utilizar el algoritmo el tiempo disminuye en un $90,00\,\%$.

5.2. Medidas de calidad de discriminación.

En esta sección comparamos la calidad de discriminación del modelo obtenido a través de la ejecución del algoritmo en R con la calidad de discriminación del modelo obtenido a través de la realización manual de los pasos de la metodología.

Comparando las tablas Tabla 5.2 y Tabla 5.3, tenemos que el estadístico KS, el índice $AUROC$, el coeficiente de $GINI$ para el modelo estimado por el algoritmo presentan un valor superior a los indicadores del modelo obtenido mediante la realización manual de la metodología.

Esto se puede justificar debido a que en algoritmo se consideran todas las variables continuas y las que se generan mediante árboles de decisión y de ellas se seleccionan las que presentan los valores más altos en las medidas de divergencia y asociación explicadas a detalle en la secciones (2.1) y (2.2) del capítulo 2, pero cuando realizamos la interacción de variables a través de árboles de decisión manualmente puede darse el caso de que obviemos ciertas variables con un alto poder predictivo, esto por el elevado número de interacciones entre variables que se deberían realizar.

i	Indicador	Valor
1	KS	0,5726
2	$AUROC$	0,8635
3	$GINI$	0,7270

Tabla 5.2: Indicadores del modelo automático.

i	Indicador	Valor
1	KS	0,5544
2	$AUROC$	0,8400
3	$GINI$	0,6800

Tabla 5.3: Indicadores del modelo manual.

Comparando ahora los gráficos de la Figura 5.1a y la Figura 5.2b, observamos que la curva ROC para el caso del modelo resultante de la ejecución del algoritmo (Figura

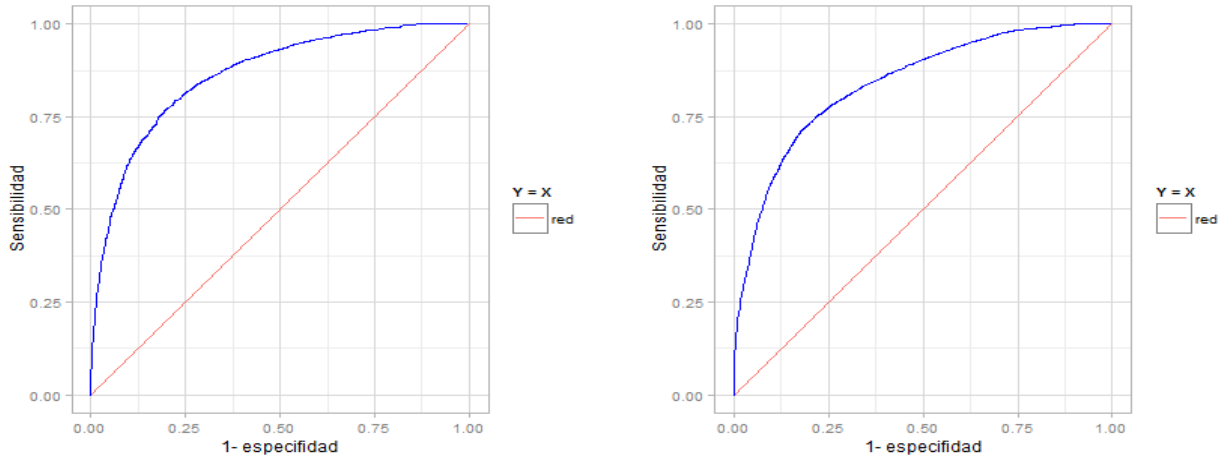

Figura 5.1: Curva ROC modelo automático-manual

5.1a) diverge en mayor medida de la recta $Y = X$, que la curva ROC del modelo resultante de la realización manual de los pasos de la metodología (Figura 5.1b).

5.3. Tabla de contingencia entre Y y $\hat{Y}$.

En esta sección comparamos el porcentaje de clasificación correcta del modelo resultante de la implementación y el modelo realizado manualmente.

El punto de corte para el modelo resultante de la implementación es de 0.601, y para el modelo realizado manualmente es de 0.698, al parecer la probabilidad de activación estimada por el modelo resultante de la implementación subestima a la probabilidad de activación estimada por el modelo realizado manualmente, esto lo podremos validar cuando analicemos las tablas de performance.

Analizando las tablas Tabla 5.4 y Tabla 5.5, tenemos que con la ejecución del algoritmo un 89,00 % y un 62,00 % de los sujetos Activa y No Activa respectivamente son clasificados correctamente. En cambio para el modelo realizado manualmente un 71,29 % y un 82,13 % de los sujetos Activa y No Activa respectivamente son clasificados correctamente; como nuestro objetivo es identificar a los sujetos Activa el modelo resultante de la implementación (Sujetos Activa clasificados correctamente: 89,00 %) sería mucho mejor que el modelo realizado manualmente (Sujetos Activa

clasificados correctamente: 71,29 %).

$Y \setminus \hat{Y}$	**Activa**	**No Activa**	% **Clasificación correcta**
Activa	8.973	1.101	89,00 %
No Activa	1.907	3.124	62,00 %

Tabla 5.4: Tabla de clasificación del modelo automático.

$Y \setminus \hat{Y}$	**Activa**	**No Activa**	% **Clasificación correcta**
Activa	7.182	2.892	71,29 %
No Activa	899	4.132	82,13 %

Tabla 5.5: Tabla de clasificación del modelo manual.

5.4. Tablas de performance.

Comparando ahora las tablas de performance del modelo resultante de la implementación (Tabla 5.6) y del modelo realizado manualmente (Tabla 5.7), observamos una diferencia significativa en los extremos de los diez intervalos de la probabilidad de activación estimada, los deciles de la probabilidad de activación estimada con el modelo resultante de la implementación son menores que los deciles de la probabilidad de activación estimada con el modelo realizado manualmente, es decir la probabilidad de activación estimada por el modelo automático subestima la probabilidad de activación estimada por el modelo manual. En lo que se refiere a la distribución de los sujetos totales, sujetos Activa, sujetos No Activa no se pueden observar diferencias significativas.

De lo anterior podemos concluir que los dos modelos resultantes presentan un alto rendimiento de discriminación a lo largo del intervalo [0,1].

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum	%Act	%No Act
0,805	0,999	1.531	10,0 %	10,0 %	1.503	14,9 %	14,9 %	28	0,6 %	0,6 %	98,2 %	1,8 %
0,619	0,805	1.493	10,0 %	20,0 %	1.425	14,1 %	29,1 %	68	1,4 %	1,9 %	95,4 %	4,6 %
0,477	0,619	1.527	10,0 %	30,0 %	1.412	14,0 %	43,1 %	115	2,3 %	4,2 %	92,5 %	7,5 %
0,360	0,477	1.505	10,0 %	40,0 %	1.325	13,2 %	56,2 %	180	3,6 %	7,8 %	88,0 %	12,0 %
0,255	0,360	1.507	10,0 %	50,0 %	1.224	12,2 %	68,4 %	283	5,6 %	13,4 %	81,2 %	18,8 %
0,157	0,255	1.503	10,0 %	60,0 %	1.068	10,6 %	79,0 %	435	8,6 %	22,0 %	71,1 %	28,9 %
0,097	0,157	1.520	10,0 %	70,0 %	862	8,6 %	87,5 %	658	13,1 %	35,1 %	56,7 %	43,3 %
0,054	0,097	1.505	10,0 %	80,0 %	645	6,4 %	93,9 %	860	17,1 %	52,2 %	42,9 %	57,1 %
0,026	0,054	1.506	10,0 %	90,0 %	432	4,3 %	98,2 %	1.074	21,3 %	73,6 %	28,7 %	71,3 %
0,010	0,026	1.508	10,0 %	100,0 %	178	1,8 %	100,0 %	1.330	26,4 %	100,0 %	11,8 %	88,2 %
Total		15.105			10.074			5.031				

Tabla 5.6: Tabla de performance del modelo automático.

Prob de activación		Sujetos totales			Sujetos Activa			Sujetos No Activa			Razón Activa:No Activa	
Min	Max	Num	%	%Acum	Num	%	%Acum	Num	%	%Acum	%Act	%No Act
0,951	0,999	1.510	10,0 %	10,0 %	1.487	14,8 %	14,8 %	23	0,5 %	0,5 %	98,5 %	1,5 %
0,920	0,951	1.511	10,0 %	20,0 %	1.412	14,0 %	28,8 %	99	2,0 %	2,4 %	93,4 %	6,6 %
0,890	0,920	1.510	10,0 %	30,0 %	1.364	13,5 %	42,3 %	146	2,9 %	5,3 %	90,3 %	9,7 %
0,846	0,890	1.511	10,0 %	40,0 %	1.324	13,1 %	55,5 %	187	3,7 %	9,0 %	87,6 %	12,4 %
0,758	0,846	1.510	10,0 %	50,0 %	1.195	11,9 %	67,3 %	315	6,3 %	15,3 %	79,1 %	20,9 %
0,609	0,758	1.511	10,0 %	60,0 %	1.020	10,1 %	77,4 %	491	9,8 %	25,1 %	67,5 %	32,5 %
0,514	0,608	1.511	10,0 %	70,0 %	809	8,0 %	85,5 %	702	14,0 %	39,0 %	53,5 %	46,5 %
0,423	0,514	1.510	10,0 %	80,0 %	698	6,9 %	92,4 %	812	16,1 %	55,2 %	46,2 %	53,8 %
0,261	0,423	1.511	10,0 %	90,0 %	578	5,7 %	98,1 %	933	18,5 %	73,7 %	38,3 %	61,7 %
0,010	0,261	1.510	10,0 %	100,0 %	187	1,9 %	100,0 %	1.323	26,3 %	100,0 %	12,4 %	87,6 %
Total		15.105			10.074			5.031				

Tabla 5.7: Tabla de performance del modelo manual.

5.5. Variables explicativas.

En esta sección se analizan las variables explicativas que conforman cada uno de los dos modelos obtenidos. En la Tabla 5.8 (resultado que genera el algoritmo en R.), tenemos los regresores del modelo obtenido tras la ejecución del algoritmo implementado en R y en la Tabla 5.9, tenemos los regresores incluidos en el modelo resultante de la realización manual de la metodología.

Podemos observar una desventaja del algoritmo generado en R, puesto que en ocasiones es necesario incluir en el modelo variables que analicen el comportamiento en los últimos 12, 36 meses anteriores al punto de observación, en este

modelo automático únicamente analizamos el comportamiento actual e histórico de 3, 6 meses anteriores al punto de observación, variables que analicen el comportamiento en otros sectores por ejemplo sector comercial, variables que analicen el comportamiento en otros tipos de crédito similares por ejemplo créditos de consumo, a pesar de que la inclusión de este tipo de variables disminuya el poder predictivo del modelo (KS, $AUROC$, $GINI$, etc.).

Finalmente procedemos a realizar la comparación entre los ajustes de los modelos manual y automático, mediante el coeficiente de determinación R^2. Luego de analizar los valores del coeficiente de determinación de cada uno de los modelos, tenemos que el modelo automático presenta un ajuste $(R^2 = 0,4779)$ superior al del modelo manual $(R^2 = 0,4779)$.

i	**Variable**	**Coeficiente**	**Error estándar**	**Significancia**	**Lim inf**	**Lim sup**
1	*Porc_Cupo_Util*	0,010	0,001	0,000	0,007	0,013
2	*Num_TC_Consumo*	0,172	0,028	0,000	0,116	0,228
3	*Max_Antiguedad_TC*	-0,001	0,001	0,009	-0,002	0,000
4	*Mejor_TC*	-0,329	0,038	0,000	-0,406	-0,253
5	*prba_TC_Abiert_Ult_3M y Copen_Vig_3_all*	2,519	0,156	0,000	2,212	2,826
6	*prba_Ln_Cupo_Prom_TC_Vig y Copen_Vig_3_all*	0,871	0,180	0,000	0,517	1,226
7	*prba_Ln_Max_Cupo_Vig y TC_Abiert_Ult_3M*	3,285	0,173	0,000	2,948	3,629
8	*d_Porc_Cupo_Util_6M y Ln_Cupo_Prom_TC_Vig*	0,831	0,073	0,000	0,687	0,974
9	*d_Cupo_Prom_TC_Consumo y Ln_Cupo_Prom_TC_Vig*	0,540	0,049	0,000	0,442	0,637
10	*d_Max_Antiguedad_TC y Cupo_Prom_TC_Vig*	0,680	0,065	0,000	0,553	0,809
11	*Constante*	-2,189	0,348	0,000	-2,873	-1,507

Tabla 5.8: Variables del modelo automático.

i	**Variable**	**Coeficiente**	**Error estándar**	**Significancia**	**Lim inf**	**Lim sup**
1	*Porc_Cupo_Util*	0,009	0,001	0,000	0,007	0,012
2	*Max_Antiguedad_TC*	-0,007	0,001	0,000	-0,008	-0,006
3	*Ln_Max_Cupo_Vig_TC*	0,120	0,012	0,000	0,096	0,144
4	*Ln_Cupo_Prom_TC_Vig*	0,187	0,008	0,000	0,171	0,203
5	*Mejor_TC*	-0,505	0,037	0,000	-0,579	-0,432
6	*rdop_3ab12_TC*	0,238	0,117	0,041	0,008	0,467
7	*rdop_3ab12_C*	0,131	0,034	0,027	-0,005	0,267
8	*Ln_rfp24_C*	0,021	0,010	0,032	0,002	0,040
9	*Saldo_Cred_Comercial_12M*	-0,172	0,061	0,005	-0,292	-0,051
10	*Ln_Saldo_Sicom*	0,021	0,005	0,000	0,012	0,031
11	*Comprometido*	-0,420	0,113	0,000	-0,641	-0,199
12	*d_Azuay*	0,315	0,085	0,000	0,148	0,482
13	*prba_Region*	3,353	0,465	0,000	2,442	4,264
14	*prba_Edad*	1,335	0,530	0,012	0,296	2,373
15	*prba_TC_Abiert_Ult_3M y Copen_Vig_3M*	4,461	0,105	0,000	4,256	4,666
16	Constante	-2,452	0,587	0,000	-3,603	-1,301

Tabla 5.9: Variables del modelo manual.

Capítulo 6

Conclusiones y recomendaciones.

En el Ecuador, luego de entrar en vigencia la resolución de la Junta Bancaria, correspondiente a la modificación en el factor de ponderación de la cuenta **6404 Créditos aprobados no desembolsados**, en la cual se incluyen los cupos de tarjetas de crédito no utilizados, para efectos del cálculo del patrimonio técnico requerido, se evidenció la necesidad de disponer de una herramienta técnica que permita predecir el hábito de consumo después de la apertura de la tarjeta.

El modelo de activación obtenido, permite estimar la probabilidad que tiene un individuo de realizar al menos un consumo en los tres siguientes meses posteriores al mes de la apertura. Constituye un aporte técnico importante para las instituciones emisoras de tarjetas de crédito, debido a que la predicción adecuada de dicha probabilidad, les permitiría identificar potenciales clientes, es decir aquellos clientes cuya probabilidad de activar la tarjeta de crédito sea elevada, para únicamente sobre ellos desarrollar las acciones correspondientes, como por ejemplo, una estrategia de marketing apropiada, para de esta manera evitar incurrir en gastos innecesarios, pues resulta menos costoso rechazar la activación de la tarjeta, que activarla y que el usuario no realice consumo alguno en una ventana de tiempo determinada.

Antes de proceder a describir los hallazgos más importantes realizados, debemos recalcar los principales aportes generados por el presente trabajo, entre ellos tenemos, la descripción detallada de cada uno de los pasos que se deben seguir para construir un modelo de activación (metodología analítica), la construcción del modelo

de activación de tarjetas de crédito, el cual, en el Ecuador resulta ser el primero de su tipo, finalmente tenemos la implementación de la metodología analítica en R, es decir la generación de un algoritmo que realiza automáticamente cada uno sus pasos con el objetivo de disminuir su tiempo de ejecución permitiendo que la computadora realice las tareas más tediosas y repetitivas.

Considerando los resultados obtenidos, podemos enumerar las siguientes conclusiones principales:

1. El valor del estadístico KS de un modelo de regresión logística tiene como cota inferior el máximo KS de las variables explicativas incluidas en la regresión, es decir, dadas $X_1, X_2, ..., X_p$, p variables explicativas con sus respectivos estadísticos $KS_{X_1}, KS_{X_2}, ..., KS_{X_p}$, tenemos que el KS del modelo final (KS_{MOD}) verifica la desigualdad (6.1).

$$KS_{MOD} \geq Max\ \{KS_{X_1}, KS_{X_2}, ..., KS_{X_p}\} \tag{6.1}$$

   lo anterior se puede justificar considerando la monotonía creciente de la función logística. Dado que el poder predictivo de una variable puede medirse mediante el KS, podemos concluir que el poder predictivo del modelo depende en gran medida del poder predictivo de cada una de sus variables explicativas.

   En nuestro caso, de las 50 variables explicativas disponibles, la variable $Copen\_Vig\_3M$ presentó un $KS = 0,3685$, este valor resulta ser el más elevado considerando los valores del KS de todas las variables del ANEXO B y el modelo final presentó un $KS_{MOD} = 0,5544$, de donde se verifica que:

$$KS_{MOD} = 0,5544 \geq Max\ \{KS_{X_1}, KS_{X_2}, ..., KS_{X_{50}}\} = 0,3685$$

2. Realizar la interacción entre variables con un elevado poder predictivo medido a través del KS, mediante árboles de decisión, permite generar una nueva variable, ya sea dummy o probabilidad de activa, con un poder predictivo superior al de las dos variables iniciales. En nuestro caso el KS de la variable $TC\_Abiert\_Ult\_3M$

es de $0,3685$ y el de la variable $Copen\_Vig\_3M$ de $0,3374$, tras la interacción mediante el árbol de decisión de la Figura 3.7, obtenemos la nueva variable $prba\_TC\_Abiert\_Ult\_3M\ y\ Copen\_Vig\_3M$, la cual presenta un KS de $0,498$, el mismo que resulta ser superior al KS de las dos variables iniciales.

3. la principal limitación del poder predictivo de un modelo de regresión, dado por los valores obtenidos para el $KS, AUROC, GINI$ de la Tabla 5.2, es la calidad de discriminación de la información histórica disponible (variables explicativas), dicha calidad puede ser evaluada mediante las medidas de divergencia y asociación presentadas en el ANEXO B.

4. Cada uno de los pasos de la metodología utilizada para la construcción del modelo de activación puede ser implementado en el lenguaje de programación R, tal como se presentó en el flujograma de la Figura 4.7 del capítulo 4. Esto es posible por su facilidad de implementación, manejo de programación orientada a objetos, y la principal, cuenta con los algoritmos más actuales y más utilizados en la generación de modelos estadísticos.

5. Comparando el estadístico KS y los índices $AUROC, GINI$ del modelo obtenido tras la realización manual de cada uno de los pasos de la metodología (modelo manual) de la Tabla 5.2, con los valores obtenidos para el modelo resultante de la ejecución del algoritmo implementado en R (modelo automático) de la Tabla 5.3 del capítulo 5, tenemos que el modelo automático tiene un mayor poder de predicción que el modelo manual. Esto se puede justificar debido a cuando se realiza la interacción entre las variables de forma manual se pueden obviar ciertas variables importantes, esto no sucede en el modelo automático, pues en este caso se realizan automáticamente todas las interacciones posibles.

La información relacionada con las transacciones y las ofertas históricas de tarjetas de crédito, resulta ser muy predictiva en la construcción de los modelos de activación, es por esta razón que es indispensable contar con procesos adecuados de recopilación y almacenamiento de dicha información, pues actualmente para las instituciones emisoras de tarjetas del Ecuador les resulta muy costoso generar las variables de la base de datos transaccional y de la base de ofertas históricas o peor aún, en

ocasiones esta información no se almacena.

Con el propósito de facilitar el trabajo futuro y mejorar la calidad de predicción de los modelos de activación de tarjetas de crédito, es necesario realizar las siguientes **recomendaciones**:

1. Recopilar y almacenar la información relacionada con transacciones y ofertas históricas de tarjetas de crédito, esto con el propósito de generar un modelo de activación con un mejor poder de discriminación y predicción.

2. Promover el uso del software libre R, en la construcción de modelos estadísticos y generación de nuevas metodologías, tanto en las instituciones educativas, como en las instituciones públicas y privadas.

3. Estudiar las posibles ventajas que implicaría la utilización de modelos de activación u otros modelos complementarios tales como modelos de deserción, lealtad, etc. en el fortalecimiento de las relaciones entre la entidad y el cliente.

Bibliografía

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

[Arnold and Emerson, 2011] Arnold, T. and Emerson, J. (2011). Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal*, 3:34–35.

[Bordeleau, 2009] Bordeleau, D. (2009). Exploring Alternative Predictive Modeling Techniques to Strengthen the Customer Relationship. *SAS Institute Inc*, page 1.

[Buckinx and Van den Poel, 2003] Buckinx, W. and Van den Poel, D. (2003). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, pages 255–256.

[Buckinx et al., 2006] Buckinx, W., Verstraeten, G., and Van den Poel, D. (2006). Predicting customer loyalty using the internal transactional database. *Expert Systems with Applications*.

[Castro, 2008] Castro, A. (2008). *Regresión Lineal*. Monografías de Matemática y Estadística, Quito.

[Engmann and Cousineau, 2011] Engmann, S. and Cousineau, D. (2011). Comparing Distributions: The Two-Sample Anderson-Darling Test as an Alternative to the Kolmogorov-Smirnoff Test. *Journal of Applied Quantitative Methods*, 6.

[Fawcett, 2005] Fawcett, T. (2005). An introduction to ROC analysis. *Institute for the Study of Learning and Expertise*, pages 361–363.

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

[Lewis, 2013] Lewis, N. (2013). *100 Statistical Tests in R*. CreateSpace Independent Publishing Platform.

[Massey, 1951] Massey, F. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, page 6878.

[Nie et al., 2011] Nie, G., Rowe, W., Zhang, L., Tian, Y., and Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*.

[Parr Rud, 2001] Parr Rud, O. (2001). *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*. John Wiley & Sons, Inc, New York.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Scholz and Stephens, 1987] Scholz, F. W. and Stephens, M. A. (1987). K-Sample Anderson-Darling Tests. *Journal of the American Statistical Association*, 82:918–919.

[Siddiqi, 2006] Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Inc, New Jersey.

ANEXO A: Descripción de variables explicativas

i	VARIABLE	DESCRIPCIÓN
1	Cant_Total_12_TC	Número de tarjetas abiertas durante los últimos 12 meses anteriores al punto de observación.
2	Cant_Total_3_TC	Número de tarjetas abiertas durante los últimos 3 meses anteriores al punto de observación.
3	Comprometido	Razón entre la cuota estimada y el ingreso que registra el sujeto al punto de observación.
4	Copen_Vig_3M	Variable binaria que toma el valor de 1 si el sujeto registra uno o más productos vigentes, abiertos durante los últimos 3 meses anteriores al punto de observación, excluyendo operaciones de tarjetas de crédito.
5	Ln_Cupo_Prom_TC_Consumo	Logaritmo natural del cupo promedio por tarjeta que registra consumos al punto de observación.
6	Deuda_Tot_12_TC	Deuda total en TC en los últimos 12 meses anteriores al punto de observación.
7	Deuda_Tot_3_TC	Deuda total en TC en los últimos 3 meses anteriores al punto de observación.
8	Dígito	Primeros dos dígitos de la cédula de ciudadanía.
9	Edad	Edad que registra el sujeto al punto de observación.
10	Ln_Cupo_Prom_TC_Vig	Logaritmo natural del cupo promedio por TC vigente al punto de observación.
11	Ln_Max_Cupo_Vig_TC	Logaritmo natural del máximo cupo de las TC vigentes al punto de observación.
12	Ln_rfp24_C	Logaritmo natural de la razón de la deuda por vencer respecto a la deuda total los últimos 24 meses anteriores al punto de observación en créditos de consumo.
13	Ln_Saldo_Sicom	Logaritmo natural del saldo de la deuda total en el sistema comercial al punto de observación.
14	Max_Antiguedad_TC	Tiempo en meses desde que el sujeto aperturó su primera TC hasta el punto de observación.
15	Mejor_TC	Mejor tarjeta que registra el sujeto al punto de observación.
16	Mold_Sicom	Tiempo en meses desde que el sujeto aperturó su primer crédito en el sector comercial hasta el punto de observación.
17	Num_Acreedores_SEPS	Número de acreedores en las instituciones reguladas por la SEPS al punto de observación.

i	VARIABLE	DESCRIPCIÓN
18	Num_Acreedores_Sicom	Número de acreedores en las instituciones del sector comercial al punto de observación.
19	Num_TC_Consumo	Número de tarjetas que registran consumos al punto de observación.
20	Num_TC_Vig	Número de tarjetas vigentes al punto de observación.
21	Porc_Acreedores_Sicom_SFR	Número de acreedores en el sector comercial sobre el número de acreedores en las instituciones reguladas por la SBS.
22	Porc_Cupo_Uti	Porcentaje de cupo utilizado al punto de observación.
23	Porc_Cupo_Uti_6M	Porcentaje de cupo utilizado en los últimos 6 meses anteriores al punto de observación.
24	Porc_Saldo_Consumo_y_TC	Suma del saldo en consumo y TC sobre la deuda total en el Sistema Crediticio Ecuatoriano al punto de observación.
25	rctot_3s12_tc	Razón entre el número de tarjetas abiertas los últimos 3 meses y el número de tarjetas abiertas los últimos 12 meses anteriores al punto de observación.
26	rctot_3s24_tc	Razón entre el número de tarjetas abiertas los últimos 3 meses y el número de tarjetas abiertas los últimos 24 meses anteriores al punto de observación.
27	rdop_3ab12_C	Razón entre entre el número de créditos de consumo los últimos 3 meses y los últimos 12 meses anteriores al punto de observación.
28	rdop_3ab12_TC	Razón entre las operaciones en TC realizadas en los últimos 3 meses y los últimos 12 meses anteriores al punto de observación.
29	rdt_12s24m_tc	Razón de la deuda total en TC en los últimos 12 meses respecto a la deuda total en los últimos 36 meses anteriores al punto de observación en la entidades reguladas por la SBS.
30	rdt_3s12_consumo	Razón de la deuda total en créditos de consumo en los últimos 3 meses respecto a la deuda total en los últimos 36 meses anteriores al punto de observación en la entidades reguladas por la SBS.
31	rdt_3s12_sfr	Razón de la deuda total en los últimos 3 meses respecto a la deuda total en los últimos 36 meses anteriores al punto de observación en la entidades reguladas por la SBS.
32	rdt_3s12_tc	Razón de la deuda total en TC en los últimos 3 meses respecto a la deuda total en los últimos 12 meses anteriores al punto de observación en la entidades reguladas por la SBS.
33	rdt_sicom_sfr	Razón de la deuda total en el sector comercial respecto a la deuda total en las entidades reguladas por la SBS al punto de observación.
34	rfp24_tc_in	Razón de la deuda por vencer respecto a la deuda total los últimos 24 meses anteriores al punto de observación en TC.
35	rope_corr_y_rot_12_tc	Razón de las operaciones corrientes y rotativas respecto al total de operaciones en TC en los últimos 12 meses anteriores al punto de observación.
36	rsal_xvencer_total_sicom	Razón entre el saldo por vencer y el saldo total de la deuda en el sector comercial al punto de observación.
37	rtc_msalcor_12	Razón del saldo vigente en operaciones de TC corrientes respecto al máximo cupo en los últimos 12 meses anteriores al punto de observación.
38	rtc_msalcor_3	Razón del saldo vigente en operaciones de TC corrientes respecto al máximo cupo en los últimos 3 meses anteriores al punto de observación.

i	VARIABLE	DESCRIPCIÓN
39	rtc_msaldifsi_3	Razón del saldo vigente en operaciones de TC diferidas respecto al máximo cupo en los últimos 3 meses anteriores al punto de observación.
40	rtc_msalrot_12	Razón del saldo vigente en operaciones de TC rotativas respecto al máximo cupo en los últimos 12 meses anteriores al punto de observación.
41	rtc_msalrot_3	Razón del saldo vigente en operaciones de TC rotativas respecto al máximo cupo en los últimos 3 meses anteriores al punto de observación.
42	rtc_opcor_12	Porcentaje de operaciones de TC corrientes respecto al total de operaciones en los últimos 12 meses anteriores al punto de observación.
43	rtc_opdifsi_12	Porcentaje de operaciones de TC diferidas respecto al total de operaciones en los últimos 12 meses anteriores al punto de observación.
44	rtc_oprot_3	Porcentaje de operaciones de TC rotativas respecto al total de operaciones en los últimos 12 meses anteriores al punto de observación.
45	rtm_3vig12_consumo	Razón entre el máximo monto vigente en créditos de consumo en los últimos 3 meses respecto al máximo monto vigente en créditos de consumo en los 12 últimos meses anteriores al punto de observación en las entidades reguladas por la SBS.
46	rtm_3vig12_sicom	Razón entre el máximo monto vigente en el sector comercial en los últimos 3 meses respecto al máximo monto vigente en el sector comercial en los 12 últimos meses anteriores al punto de observación en las entidades reguladas por la SBS.
47	Saldo_Cred_Comercial_12M	Variable binaria que toma el valor de 1 si el máximo saldo en operaciones comerciales (en casas comerciales) durante en los últimos 12 meses anteriores al punto de observación es mayor que cero.
48	Saldo_TC	Saldo total en TC en el Sistema Crediticio Ecuatoriano al punto de observación.
49	TC_Abiert_Ult_3M	Número de tarjetas abiertas los últimos 3 meses anteriores al punto de observación.
50	Tot_Por_Vencer_Sicom	Total de deuda por vencer en el sector comercial al punto de observación.
51	d_Azuay	Variable binaria que toma el valor de 1 si el sujeto pertenece a la provincia del Azuay
52	Estado_Civil	Estado civil que registra el sujeto al punto de observación.
53	Género	Género que registra el sujeto al punto de observación.
54	Mayor_Plazo_Vencido	Mayor plazo vencido que registra el sujeto en los últimos 36 meses anteriores al punto de observación en el Sistema Crediticio Ecuatoriano
55	Peor_Calf_Actual_Deuda_Directa	Peor calificación en deuda directa al punto de observación.
56	Peor_Calf_Hist_12M	Peor calificación en deuda directa en los últimos 12 meses al punto de observación.
57	Peor_Calf_Hist_36M	Peor calificación en deuda directa en los últimos 36 meses al punto de observación.
58	Peor_Calf_Seps	Peor calificación en el sector regulado por la SEPS.
59	Peor_Calf_Sicom	Peor calificación en el sector comercial al punto de observación.
60	Posee_Cred_Comercial	Variable binaria que indica si tiene o no un crédito comercial.
61	Posee_TC_Competencia	Variable binaria que indica si tiene o no una TC en la competencia.
62	Region	Región de nacimiento que registra el sujeto.

ANEXO B: Medidas de divergencia (Definición 1)

i	VARIABLE	KS	AD	|CORR|	IND_{NUM}
1	Copen_Vig_3M	0.3685	0.2014	0.3492	0.3331
2	TC_Abiert_Ult_3M	0.3374	0.2097	0.2390	0.3020
3	Ln_Cupo_Prom_TC_Vig	0.2709	0.1376	0.3693	0.2540
4	Porc_Cupo_Uti	0.2447	0.1065	0.2314	0.2157
5	Num_TC_Consumo	0.1555	0.0552	0.1902	0.1389
6	Cupo_Prom_TC_Consumo	0.1650	0.0371	0.0349	0.1264
7	Ln_Max_Cupo_Vig_TC	0.1349	0.0286	0.2547	0.1256
8	Porc_Cupo_Uti_6M	0.1501	0.0312	0.0190	0.1132
9	Max_Antiguedad_TC	0.1308	0.0271	0.0936	0.1063
10	Ln_Saldo_Sicom	0.1132	0.0246	0.1018	0.0943
11	Comprometido	0.1029	0.0183	0.0674	0.0824
12	Tot_Por_Vencer_Sicom	0.1004	0.0204	0.0293	0.0772
13	Mejor_TC	0.0811	0.0213	0.1016	0.0711
14	rsal_xvencer_total_sicom	0.0819	0.0121	0.0737	0.0671
15	Porc_Acreedores_Sicom_SFR	0.0894	0.0117	0.0171	0.0666
16	Num_TC_Vig	0.0833	0.0130	0.0342	0.0643
17	Edad	0.0763	0.0108	0.0767	0.0632
18	rdt_sicom_sfr	0.0698	0.0043	0.0158	0.0513
19	Num_Acreedores_Sicom	0.0544	0.0065	0.0600	0.0453
20	Dígito	0.0522	0.0031	0.0188	0.0390
21	Cant_Total_3_TC	0.0339	0.0012	0.0268	0.0266
22	Cant_Total_12_TC	0.0326	0.0013	0.0268	0.0257
23	Num_Acreedores_SEPS	0.0244	0.0060	0.0487	0.0231
24	rdt_12s24m_tc	0.0268	0.0005	0.0141	0.0202
25	rtc_opdifsi_12	0.0259	0.0012	0.0157	0.0199
26	Deuda_Tot_3_TC	0.0250	0.0007	0.0073	0.0183
27	Deuda_Tot_12_TC	0.0249	0.0006	0.0062	0.0181
28	rtc_msaldifsi_3	0.0236	0.0007	0.0125	0.0179
29	Porc_Saldo_Consumo_y_TC	0.0201	0.0004	0.0178	0.0159
30	rctot_3s24_tc	0.0209	0.0003	0.0114	0.0158
31	Mold_Sicom	0.0191	0.0002	0.0123	0.0146

i	VARIABLE	KS	AD	|CORR|	IND_{NUM}
32	rdt_3s12_tc	0.0181	0.0003	0.0082	0.0135
33	rtc_msalrot_3	0.0181	0.0003	0.0058	0.0133
34	Saldo_TC	0.0181	0.0003	0.0002	0.0127
35	rope_corr_y_rot_12_tc	0.0179	0.0002	0.0003	0.0126
36	rtm_3vig12_consumo	0.0163	0.0000	0.0096	0.0123
37	rdt_3s12_sfr	0.0168	0.0003	0.0054	0.0123
38	rdt_3s12_consumo	0.0168	0.0002	0.0028	0.0120
39	rtc_msalcor_3	0.0143	0.0005	0.0178	0.0118
40	rfp24_tc_in	0.0152	0.0003	0.0116	0.0118
41	rtc_msalcor_12	0.0134	0.0003	0.0180	0.0112
42	rtc_msalrot_12	0.0134	0.0002	0.0046	0.0098
43	rtc_oprot_3	0.0132	0.0003	0.0018	0.0094
44	rctot_3s12_tc	0.0116	0.0001	0.0095	0.0090
45	rdop_3ab12_C	0.0107	0.0002	0.0125	0.0087
46	rtc_opcor_12	0.0118	0.0003	0.0025	0.0085
47	Saldo_Cred_Comercial_12M	0.0099	0.0005	0.0135	0.0083
48	rdop_3ab12_tc	0.0057	0.0005	0.0136	0.0054
49	Ln_rfp24_C	0.0065	0.0000	0.0040	0.0049
50	rtm_3vig12_sicom	0.0023	0.0000	0.0019	0.0018

Medidas de asociación (Definición 1).

i	VARIABLE	$CONT$	DIC	VI	IND_{CAT}
1	Peor_Calf_Sicom	0.1590	0.0759	0.1302	0.1395
2	Peor_Calf_Actual_Deuda_Directa	0.1436	0.0684	0.0919	0.1233
3	Peor_Calf_Hist_12M	0.1248	0.0593	0.0793	0.1071
4	Peor_Calf_Hist_36M	0.1187	0.0566	0.0922	0.1036
5	Estado_Civil	0.1210	0.0575	0.0665	0.1028
6	Mayor_Plazo_Vencido	0.1143	0.0542	0.0455	0.0954
7	Region	0.1047	0.0496	0.0507	0.0882
8	Peor_Calf_Seps	0.0737	0.0349	0.0306	0.0616
9	Género	0.0470	0.0223	0.0100	0.0383
10	Posee_Cred_Comercial	0.0187	0.0090	0.0017	0.0150
11	Posee_TC_Competencia	0.0172	0.0082	0.0014	0.0138
12	d_Azuay	0.0030	0.0016	0.0000	0.0024

ANEXO C: Medidas de divergencia (Definición 2)

i	VARIABLE	KS	AD	|CORR|	IND_{NUM}
1	Ln_Cupo_Prom_TC_Vig	0.2779	0.1382	0.3684	0.3050
2	copen_vig_3M	0.2935	0.1434	0.2852	0.2910
3	TC_Abiert_Ult_3M	0.2672	0.1505	0.2027	0.2478
4	Ln_Max_Cupo_Vig_TC	0.1526	0.0310	0.2787	0.1904
5	Porc_Cupo_Uti	0.1994	0.0595	0.1497	0.1844
6	Num_TC_Consumo	0.1710	0.0675	0.2078	0.1820
7	Ln_Saldo_Sicom	0.1227	0.0284	0.1111	0.1192
8	Cupo_Prom_TC_Consumo	0.1443	0.0296	0.0229	0.1078
9	rsal_xvencer_total_sicom	0.1014	0.0177	0.0938	0.0991
10	Comprometido	0.1055	0.0184	0.0704	0.0949
11	Num_TC_Vig	0.0909	0.0161	0.0953	0.0922
12	Tot_Por_Vencer_Sicom 0	0.1185	0.0279	0.0304	0.0920
13	Porc_Cupo_Uti_6M	0.1267	0.0247	0.0067	0.0907
14	Num_Acreedores_Sicom	0.0735	0.0116	0.0814	0.0758
15	Porc_Acreedores_Sicom_SFR	0.1008	0.0144	0.0158	0.0753
16	Max_Antiguedad_TC	0.0632	0.0035	0.0440	0.0574
17	Edad	0.0502	0.0046	0.0529	0.0510
18	Mejor_TC	0.0490	0.0131	0.0452	0.0478
19	rdt_sicom_sfr	0.0601	0.0060	0.0145	0.0464
20	Digito	0.0444	0.0029	0.0123	0.0347
21	Num_Acreedores_SEPS	0.0202	0.0042	0.0426	0.0269
22	rdt_3s12_consumo	0.0263	0.0006	0.0146	0.0227
23	rtc_opdifsi_12	0.0217	0.0008	0.0142	0.0194
24	rdop_3ab12_C	0.0158	0.0011	0.0225	0.0178
25	rfp24_tc_in	0.0182	0.0003	0.0148	0.0171
26	rdt_12s24m_tc	0.0215	0.0002	0.0064	0.0169
27	rtc_msaldifsi_3	0.0210	0.0004	0.0074	0.0169
28	Ln_rfp24_C	0.0157	0.0003	0.0134	0.0150
29	rdt_3s12_sfr	0.0167	0.0002	0.0075	0.0139
30	Deuda_Tot_12_TC	0.0178	0.0003	0.0044	0.0137
31	Deuda_Tot_3_TC	0.0170	0.0003	0.0060	0.0137

i	VARIABLE	KS	AD	|CORR|	IND_{NUM}
32	Mold_Sicom	0.0161	0.0001	0.0066	0.0132
33	rope_corr_y_rot_12_tc	0.0173	0.0002	0.0032	0.0130
34	rdt_3s12_tc	0.0144	0.0001	0.0066	0.0120
35	Porc_Saldo_Consumo_y_TC	0.0129	0.0001	0.0090	0.0117
36	rtc_msalrot_12	0.0148	0.0001	0.0028	0.0112
37	Saldo_TC	0.0147	0.0002	0.0014	0.0107
38	Saldo_Cred_Comercial_12M	0.0092	0.0004	0.0126	0.0102
39	rctot_3s24_tc	0.0130	0.0001	0.0014	0.0095
40	rtc_oprot_3	0.0127	0.0003	0.0017	0.0094
41	rtc_msalrot_3	0.0118	0.0001	0.0035	0.0093
42	rtm_3vig12_consumo	0.0109	0.0001	0.0053	0.0092
43	rtc_msalcor_3	0.0108	0.0002	0.0024	0.0082
44	rdop_3ab12_tc	0.0061	0.0006	0.0128	0.0081
45	Cant_Tot_3_TC	0.0102	0.0001	0.0023	0.0078
46	Cant_Tot_12_TC	0.0092	0.0001	0.0022	0.0071
47	rtc_msalcor_12	0.0077	0.0000	0.0038	0.0065
48	rctot_3s12_tc	0.0063	0.0000	0.0051	0.0059
49	rtc_opcor_12	0.0074	0.0001	0.0024	0.0059
50	rtm_3vig12_sicom	0.0014	0.0000	0.0004	0.0011

Medidas de asociación (Definición 2).

i	VARIABLE	$CONT$	DIC	VI	IND_{CAT}
1	Peor_Calf_Sicom	0.1399	0.0671	0.0922	0.1132
2	Peor_Calf_Actual_Deuda_Directa	0.1235	0.0591	0.0654	0.0983
3	Peor_Calf_Hist_12M	0.1029	0.0491	0.0469	0.0811
4	Peor_Calf_Hist_36M	0.0842	0.0405	0.0346	0.0661
5	Mayor_Plazo_Vencido	0.0844	0.0406	0.0226	0.0650
6	Region	0.0624	0.043	0.0172	0.0608
7	Estado_Civil	0.0610	0.0290	0.0163	0.0469
8	Peor_Calf_SEPS	0.0583	0.0277	0.0169	0.0449
9	Genero	0.0351	0.0167	0.0052	0.0265
10	d_Azuay	0.0020	0.0086	0.0003	0.0225
11	Posee_TC_Competencia	0.0148	0.0071	0.0009	0.0111
12	Posee_Cred_Comercial	0.0116	0.0057	0.0006	0.0087

ANEXO D: Algoritmo implementado en R

```
#####################################################################
####                      R-SCORE                                ####
####    EPN                                                      ####
####    AUTOR: ALEX PÉREZ TATAMUÉS                               ####
#####################################################################
####   ELIMINANDO HISTORIAL RDATA
rm(list=ls())
#####################################################################
####               LIBRERIAS NECESARIAS                       #####
require(foreign)                                                #####
require(party)                                                  #####
require(dgof)                                                   #####
require(kSamples)                                               #####
require(gWidgets)                                               #####
require(gWidgetstcltk)                                          #####
options(guiToolkit="tcltk")                                     #####
require(ggplot2)                                                #####
#####################################################################

#####################################################################
####          CARGANDO BASE ORIGINAL DE TRABAJO.             ####
#####################################################################
####  DIRECTORIO DE TRABAJO
```

```
    msn <- paste("Ingrese dirección de la base","\n",
                  "Ejm: C:/Users/Documents/Rscore modelo1")
    dir <- ginput(message=msn,title = "DIRECCIÓN BASE DE TRABAJO",
                                                icon = c("info"))
    dir <- as.character(dir)
    setwd(dir)
####  NOMBRE DE LA BASE DE DATOS
  list.files()
  msn <- paste("Ingrese nombre de BDD","\n","Ejm: base_arboles.sav")
  nombre_base <- ginput(message=msn,title = "NOMBRE BDD",icon = c("info"))
     nombre_base<- as.character(nombre_base)
      BDD_original <-read.spss(file=nombre_base, use.value.labels=FALSE,
                               to.data.frame=TRUE)
   BDD <- BDD_original
   dim(BDD)
####  FILTROS INICIALES
   nom <- as.character(names(BDD))
#### BANCARIZADOS=1
    col <- grep("BANCARIZADOS",nom)
    BDD<-subset(BDD,BDD[,col]==1)
    BDD<-BDD[,-col]
    nom <- as.character(names(BDD))
#### ANT_3M=0
    col <- grep("ANT_3M",nom)
    BDD<-subset(BDD,BDD[,col]==0)
    BDD<-BDD[,-col]
    nom <- as.character(names(BDD))
##### NOMBRE DE LA VARIABLE DEPENDIENTE
    msn <- "Ingrese nombre de Var Dep. Ejm: GB_30"
    nom_gb <- ginput(message=msn,title = "NOMBRE VAR DEP",
                     icon = c("info"))
    nom_gb<- as.character(nom_gb)
```

```
col<- grep(nom_gb,nom)[1]
BDD<-subset(BDD,BDD[,col]<=1)
GB <- BDD[,col]
BDD<-BDD[,-col]
table(GB)
nom <- as.character(names(BDD))
dim(BDD)
#####################################################################
####                FILTRADO VARIABLES                           ####
#####################################################################
#####   FUNCION PARA MONOTONIA DE TASA DE MALO RESPECTO A TASA DE BUENO
monotona <- function(GB,var){
   bivar <- data.frame(GB,var)
 #ordenamos la variable y creamos los puntos de corte (num de fila)
  data_ord <- bivar[order(bivar[,2],decreasing=FALSE),]
  h <- floor(dim(data_ord)[1]/20)
  frac <- seq(h,20*h,h)
 # Tasas de buenos y malos por rango de la variable
  pmb <- c(0,0)
  for (i in seq(1:(length(frac)-1))){
    m <- sum(data_ord[1:frac[i],1])
    pm <- m/ sum(data_ord[,1])
    b <- (frac[i]- m)
    pb <- b/(nrow(data_ord)-sum(data_ord[,1]))
    pmb <- rbind(pmb,c(pm,pb))
 }
  pmb <- rbind(pmb[-1,],rep(1,2))
 # Condicion para monotonia: tasa malo <= tasa bueno
  cond1 <- all(pmb[,1] <= pmb[,2])
 # Condicion para monotonia: tasa malo >= tasa bueno
  cond2 <- all(pmb[,1] >= pmb[,2])
 # Condicion para monotonia: tasa m < tasa b o tasa m > tasa b
```

```
  if(cond1 | cond2){
    monotona <- TRUE
  }else{
    monotona <- FALSE
  }
  return(monotona)
}
#######   FUNCION ESTADISTICO AD
ad.test_modif <- function(x1,x2){
  na.remove <- function(x) {
    na.status <- lapply(x, is.na)
    k <- length(x)
    x.new <- list()
    na.total <- 0
    for (i in 1:k) {
      x.new[[i]] <- x[[i]][!na.status[[i]]]
      na.total <- na.total + sum(na.status[[i]])
    }
    list(x.new = x.new, na.total = na.total)
  }
  samples <- list(x1,x2)
  out <- na.remove(samples)
  na.t <- out$na.total
  samples <- out$x.new
  k <- length(samples)
  ns <- sapply(samples, length)
  if (any(ns == 0))
    stop("One or more samples have no observations.")
  x <- NULL
  for (i in 1:k) x <- c(x, samples[[i]])
  n <- length(x)
  Z.star <- sort(unique(x))
```

```
  ad <- .Call("doAdkTestStat", as.integer(k), as.double(x),
              as.integer(ns), as.double(Z.star))
  ad <- ad/min(ns)
  return(ad[1])
}
###    FUNCION: INDICADORES PARA VAR NUMERICAS
ind_var_num <-function (GB,variable){
 ##variable <- BDD[,5]
  vars <- data.frame(GB,variable)
  vars_m <- subset(vars,subset=vars[,1]==1)
  vars_b <- subset(vars,subset=vars[,1]==0)
 ## Estadistico KS
  ks <- dgof::ks.test(vars_m[,2],vars_b[,2],alternative="two.sided")
  ks <- round(as.numeric(ks$statistic),4)
 ## Estadistico AD
  if(ks>0){
    AD <- round(ad.test_modif(vars_m[,2],vars_b[,2]),4)
  }else{
    AD <- 0}
 ## Coef de correlación de pearson
  corr <- round(abs(as.numeric(cor.test(vars[,1],vars[,2],method="pearson",
         conf.level = 0.95,na.action=getOption("na.action"))[4])),4)
  return(c(ks,AD,corr))
}
###  FUNCION: ind_var_cat INDICADORES PARA VAR CATEGORICAS
 ## DIC
  RMR <- function(x,y,pesos=NULL){
  if(!is.null(pesos)){
    rmr <- sqrt(sum((pesos/sum(pesos))*((x-y)^2)))
  }else{
    rmr <- sqrt(sum(((x-y)^2))/length(x))
  }
```

```
  return(rmr)
}
### VI (IV)  x:Activa=1 - y:No Activa=0
VI <- function(x,y){
  aux <- ifelse(x/sum(x)==0,1,x/sum(x))
  wof <- log((y/sum(y))/aux)
  wof <- ifelse(wof==-Inf,0,wof)
  VI <-   sum(((y/sum(y))-(x/sum(x)))*wof)
  return(VI)
}
ind_var_cat <- function (GB,variable){
  vars <- data.frame(GB,variable)
 ### PRUEBA INDEPENDENCIA - COEF CONTINGENCIA
  chisq<-chisq.test(GB,variable)
  chisq<- as.numeric(chisq$statistic)
  cont <- round(sqrt(chisq/(chisq+length(variable))),4)
 ### ODDS VS % INICIAL DE MALO
  tc <- table(GB,variable)
  tot <- ifelse ((tc[1,] + tc[2,])==0,1,(tc[1,] + tc[2,]))
  odds <- tc[2,]/tot
  p <- as.vector (table(GB))
  porc_m <- rep(p[2]/sum(p),dim(tc)[2])
  rmr <- round(RMR(odds,porc_m,pesos=tot),4)
 ### VALOR DE INFORMACIÓN
  vi <- round(VI(tc[1,],tc[2,]),4)
  return(c(cont,rmr,vi))
}
#####################################################################
####      LLAMADO FUNCIONES FILTRADO VARIABLES                  ####
#####################################################################
#### Se calculan indicadores para var num y cat
  indicador_num <- numeric(4)
```

```
  indicador_cat <- numeric(4)
  sin_info <- numeric(3)
 for(i in seq(1:ncol(BDD))){
   no_missing <- length(BDD[!is.na(BDD[,i]),i])/length (BDD[,i])
   if(no_missing > 0.5){
    if(is.numeric(BDD[,i])){
      ind <-c(i,names(BDD)[i],monotona(GB,BDD[,i])*1,
                                        ind_var_num (GB,BDD[,i]))
      indicador_num <- rbind(indicador_num,ind)
    }
    if(is.factor(BDD[,i]) & (nlevels(BDD[,i])>=2)){
      ind <-c(i,names(BDD)[i],ind_var_cat (GB,BDD[,i]))
      indicador_cat <- rbind(indicador_cat,ind)
    }
  }else{
    ind <-c(i,names(BDD)[i],"Info menor 50%")
    sin_info <- rbind(sin_info,ind)
  }
}
### INDICADORES VARIABLES NUMERICAS
   indicador_num <- rbind(indicador_num[-1,])
   indicador_num <- data.frame (as.numeric (indicador_num[,1]),
                    indicador_num[,2],
                    as.numeric (indicador_num[,3]),
                    as.numeric (indicador_num[,4]),
                    as.numeric (indicador_num[,5]),
                    as.numeric (indicador_num[,6]))
 ### INDICADOR TOTAL
  IND <- 0.70*indicador_num[,4]+0.20*indicador_num[,5]+
         0.10*indicador_num[,6]
  indicador_num <- data.frame (indicador_num,IND)
  colnames(indicador_num) <- c("COL_VARIABLE","VARIABLE","Monotona",
```

```
                                  "KS","AD","ABS_CORR","IND")
  indicador_num <- indicador_num[order(indicador_num[,7],decreasing=TRUE),]
### INDICADORES PARA VARIABLES CATEGÓRICAS
  indicador_cat <- rbind(indicador_cat[-1,])
  indicador_cat <- data.frame (as.numeric (indicador_cat[,1]),
                       indicador_cat[,2],
                       as.numeric (indicador_cat[,3]),
                       as.numeric (indicador_cat[,4]),
                       as.numeric (indicador_cat[,5]))
### INDICADOR TOTAL
  IND <- 0.70*indicador_cat[,3]+0.20*indicador_cat[,4]+
                  0.10*indicador_cat[,5]
  indicador_cat <- data.frame (indicador_cat,IND)
  colnames(indicador_cat) <- c("COL_VARIABLE","VARIABLE",
                                  "CONT","DIC","VI","IND")
  indicador_cat <- indicador_cat[order(indicador_cat[,6],
                                       decreasing=TRUE),]
View(indicador_num)
View(indicador_cat)
View(sin_info)
write.table(file="indicador_num.csv",indicador_num,dec=",",sep="\t")
write.table(file="indicador_cat.csv",indicador_cat,dec=",",sep="\t")
write.table(file="sin_info.csv",sin_info,dec=",",sep="\t")
#################################################################
####          SELECCION DE VARIABLES CONTINUAS               ####
#################################################################
###  Numero de variables continuas a seleccionar
  msn <- paste("Ingrese el número de variables continuas a seleccionar"
                   ,"\n","Ejm: 5")
  n_cont <- ginput(message=msn,title = "SELECCIÓN VAR CONTINUAS",
                icon = c("info"))
  n_cont <- as.numeric(n_cont)
```

```
  index_var <- subset(indicador_num,indicador_num$IND>0)[,1]
  cont_var <- c(0)
 for(k in 1:length(index_var)){
   cuts <- quantile(BDD[,index_var[k]],seq(0.1,0.9,0.1),na.rm=TRUE)
   n_cut <- length(cuts)
   if(!all(cuts[1:(n_cut-2)]==cuts[2:(n_cut-1)])){
     cont_var <- c(cont_var,index_var[k])
    }else{
     cont_var<- c(cont_var,-999)
    }
  }
cont_var <- cont_var[cont_var>0]
var_continuas <- BDD[,cont_var][,1:20]
write.table(file="VARIABLES CONTINUAS.csv",var_continuas,dec=",",sep="\t")
View(var_continuas)
#####################################################################
####       FUNCIONES ARBOLES DE DECISION VAR NUMERICAS       ####
#####################################################################
### GRAFICO DE SEDIMENTACIÓN: ELECCION DEL NUM DE VARIABLES A CORTAR
graf_sedm <- function(x,titulo=NULL){
  if(!is.null(titulo)){
    INDICADOR <- x[order(x,decreasing=TRUE)]
    VARIABLES <- c(1:length(x))
    GRAF <- data.frame(INDICADOR,VARIABLES)
    g <- ggplot(data=GRAF, aes(x=VARIABLES,y=INDICADOR))
    g + geom_point(colour="brown",shape=20,alpha=1,size=2.7)+
      geom_line(colour="brown",size=0.7)+labs(title=as.character(titulo))+
      theme_light()
        }
}
#####################################################################
####               ARBOLES DE UNA VARIABLE                    ####
```

```
#####################################################################
#### Graficar Arbol de una variable
 graf_arbol1v <- function(GB,BDD,indice,n,tipo="numerica"){
  archivo <- paste("Arbol_1_var_",tipo,".pdf",sep="")
  pdf(file = archivo,width=17,height=7)
  n_var <- indice [(1:n)]
  GBa <- ifelse(GB==1,"MALO","BUENO")
  d <- data.frame(GBa,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  #y <- d[,1]
  for (i in seq(1:length(n_var))){
    #arb cambia si x1 es definida en .GlobalEnv
    #n_var[i]+1 pues d <- GB,BDD se traslada una col
    assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
    #x1 <- d[,n_var[i]]
    arb<-ctree(y~x1,data=d)
    plot(arb,main =paste("x1 =",names(BDD)[n_var[i]]))
  }
  dev.off()
}
 #### Calcular prbm, dummys arbol 1 variable
 dummys_arbol1v <- function(GB,BDD,indice,n){
  n_var <- indice [(1:n)]
  GBa <- ifelse(GB==1,"MALO","BUENO")
  d <- data.frame(GBa,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  #y <- d[,1]
 ### PRBM_DUM_1VAR sera la base de la prbm y variables binarias
  PRBM_DUM_1VAR <- GB
  nomb <- c("GB")
 ### porc_m es % de MALOS en la base de modelamiento
 ### (Se utiliza para calcular var binarias)
```

```
 p <- as.vector (table(GB))
 porc_m <- p[2]/sum(p)
 for (i in seq(1:n)){
   assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
   arb<-ctree(y~x1,data=d)
 ####  Tabla de pesos de nodos
   pesos<- as.data.frame(table(where(arb),GB))
   nod <- dim(pesos)[1]/2
   prbm<- numeric(nod)
   for(k in seq(1:nod)){
     prbm[k] <- round(pesos[k+nod,3]/(pesos[k,3]+pesos[k+nod,3]) ,2)
   }
   pesos <- data.frame(pesos[(1:nod),1],prbm)
   colnames(pesos)<-c("Nodo","prbm")
 ####   Se crean las variables prbm
   w <-ifelse(where(arb) %in% pesos[1,1],pesos[1,2],0)
   n0=2
   while(n0<=nod){
     v <- ifelse(where(arb) %in% pesos[n0,1],pesos[n0,2],0)
     w <- w+v
     n0 <- n0+1
   }
 ####   Se crean las variables binarias (dummys)
   dum <- ifelse(w>=porc_m,1,0)
 ####  Base con variables probm y dum
  PRBM_DUM_1VAR <- data.frame(PRBM_DUM_1VAR,w,dum)
  nom <- c(paste("prbm",names(BDD)[n_var[i]],sep="_"),
           paste("d",names(BDD)[n_var[i]],sep="_"))
  nomb <- c(nomb,nom)
}
 colnames(PRBM_DUM_1VAR) <- nomb
return(PRBM_DUM_1VAR)
```

```
}
#####################################################################
####                ARBOLES DE DOS VARIABLES                     ####
#####################################################################
#### Graficar Arbol de dos variables
graf_arbol2v <- function(GB,BDD,indice,n,tipo="numerica"){
  archivo <- paste("Arbol_2_var_",tipo,"s.pdf",sep="")
  pdf(file = archivo,width=17,height=7)
  n_var <- indice [(1:n)]
  GBa <- ifelse(GB==1,"MALO","BUENO")
  d <- data.frame(GBa,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  #y <- d[,1]
  for (i in seq(1:n)){
    assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
    #x1 <- d[,n_var[i]]
    for(j in seq(1:min(i,cte))){
    if(j!=i){
    assign("x2",d[,(n_var[j]+1)],envir=.GlobalEnv)
    #x2 <- d[,n_var[j]]
    arb<-ctree(y~x1+x2,data=d)
    plot(arb,main =paste("x1 =",names(BDD)[n_var[i]],"y" ,"x2 =",
                names(BDD)[n_var[j]] ))
    }
     next
  }
}
  dev.off()
}
#### Calcular prbm, dummys arbol 2 variables
 dummys_arbol2v <- function(GB,BDD,indice,n){
  n_var <- indice [(1:n)]
```

```
 GBa <- ifelse(GB==1,"MALO","BUENO")
 d <- data.frame(GBa,BDD)
 assign("y",d[,1],envir=.GlobalEnv)
 #y <- d[,1]
 ### PRBM_DUM_2VAR sera la base de la prbm y variables binarias
 PRBM_DUM_2VAR <- GB
 nomb <- c("GB")
 p <- as.vector (table(GB))
 porc_m <- p[2]/sum(p)
 for (i in seq(1:n)){
   assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
   #x1 <- d[,n_var[i]]
   for(j in seq(1:min(i,cte))){
   if(j!=i){
   assign("x2",d[,(n_var[j]+1)],envir=.GlobalEnv)
   #x2 <- d[,n_var[j]]
   arb<-ctree(y~x1+x2,data=d)
####  Tabla de pesos de nodos
  pesos<- as.data.frame(table(where(arb),GB))
  nod <- dim(pesos)[1]/2
  prbm<- numeric(nod)
  for(k in seq(1:nod)) {
     prbm[k] <- round(pesos[k+nod,3]/(pesos[k,3]+pesos[k+nod,3]) ,2)
   }
  pesos <- data.frame(pesos[(1:nod),1],prbm)
  colnames(pesos)<-c("Nodo","prbm")
####   Se crean las variables prbm=w
  w <-ifelse(where(arb) %in% pesos[1,1],pesos[1,2],0)
  n0 <- 2
  while(n0<=nod){
    v <- ifelse(where(arb) %in% pesos[n0,1],pesos[n0,2],0)
    w <- w+v
```

```
     n0 <- n0+1
    }
 ####   Se crean las variables binarias  (dummys)
  dum <- ifelse(w>=porc_m,1,0)
 ####  Base con variables probm y dum
  PRBM_DUM_2VAR <- data.frame(PRBM_DUM_2VAR,w,dum)
  nom <- c(paste("prbm",names(BDD)[n_var[i]],names(BDD)[n_var[j]],sep="_"),
       paste("d",names(BDD)[n_var[i]],names(BDD)[n_var[j]],sep="_"))
      nomb <- c(nomb,nom)
    }
     next
 }
}
colnames(PRBM_DUM_2VAR) <- nomb
 return(PRBM_DUM_2VAR)
}
#####################################################################
####                ARBOLES DE TRES VARIABLES                    ####
#####################################################################
#### Graficar Arbol de tres variables
graf_arbol3v <- function(GB,BDD,indice,n,tipo="numerica"){
  archivo <- paste("Arbol_3_var_",tipo,"s.pdf",sep="")
  pdf(file = archivo,width=17,height=7)
  n_var <- indice [(1:n)]
  GBa <- ifelse(GB==1,"MALO","BUENO")
  d <- data.frame(GBa,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  #y <- d[,1]
  for (i in seq(1:n)){
    assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
    #x1 <- d[,n_var[i]]
  for(j in seq(1:i)){
```

```
   if(j!=i){
    assign("x2",d[,(n_var[j]+1)],envir=.GlobalEnv)
    #x2 <- d[,n_var[j]]
    for(k in seq(1:n)){
      if(k!=i & k!=j){
      assign("x3",d[,(n_var[k]+1)],envir=.GlobalEnv)
      #x3 <- d[,n_var[k]]
      arb<-ctree(y~x1+x2+x3,data=d)
      plot(arb,main =paste("x1 =",names(BDD)[n_var[i]],
                   "y" ,"x2 =",names(BDD)[n_var[j]],
                   "y" ,"x3 =",names(BDD)[n_var[k]]))
      }
      next
    }
   }
   next
  }
 }
dev.off()
}
####  Calcular prbm, dummys arbol de 3 variables
dummys_arbol3v <- function(GB,BDD,indice,n){
  n_var <- indice [(1:n)]
  GBa <- ifelse(GB==1,"MALO","BUENO")
  d <- data.frame(GBa,BDD)
  assign("y",d[,1],envir=.GlobalEnv)
  #y <- d[,1]
  PRBM_DUM_3VAR_CAT <- GB
  nomb <- c("GB")
  p <- as.vector (table(GB))
  porc_m <- p[2]/sum(p)
  for (i in seq(1:n)){
```

```
  assign("x1",d[,(n_var[i]+1)],envir=.GlobalEnv)
  #x1 <- d[,n_var[i]]
 for(j in seq(1:i)){
  if(j!=i){
  assign("x2",d[,(n_var[j]+1)],envir=.GlobalEnv)
  #x2 <- d[,n_var[j]]
   for(h in seq(1:n)){
   if(h!=i & h!=j){
   assign("x3",d[,(n_var[h]+1)],envir=.GlobalEnv)
   #x3 <- d[,n_var[h]]
   arb<-ctree(y~x1+x2+x3,data=d)
   ####  Tabla de pesos de nodos
   pesos<- as.data.frame(table(where(arb),GB))
   nod <- dim(pesos)[1]/2
   prbm<- numeric(nod)
   for(k in seq(1:nod)) {
   prbm[k] <- round(pesos[k+nod,3]/(pesos[k,3]+pesos[k+nod,3]) ,2)
   }
   pesos <- data.frame(pesos[(1:nod),1],prbm)
   colnames(pesos)<-c("Nodo","prbm")
   ####   Se crean las variables prbm=w
   w <-ifelse(where(arb) %in% pesos[1,1],pesos[1,2],0)
   n0 <- 2
   while(n0<=nod){
   v <- ifelse(where(arb) %in% pesos[n0,1],pesos[n0,2],0)
   w <- w+v
   n0 <- n0+1
 }
####   Se crean las variables binarias  (dummys)
dum <- ifelse(w>=porc_m,1,0)
####  Base con variables probm y dum
PRBM_DUM_3VAR_CAT <- data.frame(PRBM_DUM_3VAR_CAT,w,dum)
```

```
  nom <- c(paste("prbm",names(BDD)[n_var[i]],names(BDD)[n_var[j]],
  names(BDD)[n_var[h]],sep="_"),
  paste("d",names(BDD)[n_var[i]],names(BDD)[n_var[j]],
  names(BDD)[n_var[h]],sep="_"))
        nomb <- c(nomb,nom)
      }
      next
     }
    }
   next
  }
 }
colnames(PRBM_DUM_3VAR_CAT) <- nomb
return(PRBM_DUM_3VAR_CAT)
}
#####################################################################
####         LLAMADO FUNCIONES ARBOLES DE DECISION           ####
#####################################################################
### GRAFICO DE SEDIMENTACIÓN PARA ELEGIR EL NUM DE VAR A CORTAR
graf_sedm(indicador_num[,7],titulo="IND NUM")
print("MEDIANTE EL GRAFICO ELIJA EL NUMERO DE VARIABLES A CORTAR")
#### GRAFICO ARBOLES DE DECISION 1 VARIABLE NUMERICA
msn <- paste("Ingrese el número n de variables para las cuales","\n",
           "se graficarán los arboles de decisión")
n <- ginput(message=msn,title = "ARBOLES DE UNA VARIABLE",
          icon = c("info"))
n <- as.numeric(n)
graf_arbol1v (GB,BDD,indice=indicador_num[,1],n,tipo="numerica")
#### CALCULO PRBM, DUMMYS ARBOL 1 VARIABLE NUMERICA
msn <- paste("Ingrese el número n de variables para las cuales","\n",
           "se calcularán las variables dummys")
n <- ginput(message=msn,title = "DUMMYS ARBOLES DE UNA VARIABLE",
```

```
          icon = c("info"))
n <- as.numeric(n)
PRBM_DUM_1VAR <- dummys_arbol1v (GB,BDD,indice=indicador_num[,1],n)
View(PRBM_DUM_1VAR)
#### GRAFICO ARBOLES DE DECISION 2 VARIABLES NUMERICAS
  msn <- paste("Ingrese el número n de variables para las cuales","\n",
               "se graficarán los arboles de decisión")
  n <- ginput(message=msn,title = "ARBOLES DE DOS VARIABLES",
              icon = c("info"))
 n <- as.numeric(n)
#### Número de variables a interaccionar
  msn <- paste("Ingrese el número n de variables a interaccionar","\n",
               "Ejemplo: 5")
  cte <- ginput(message=msn,title = "VARIABLES A INTERACCIONAR",
                icon = c("info"))
  cte <- as.numeric(cte)
  graf_arbol2v (GB,BDD,indice=indicador_num[,1],n,tipo="numerica")
#### CALCULO PRBM, DUMMYS ARBOL 2 VARIABLES NUMERICAS
  msn <- paste("Ingrese el número n de variables para las cuales","\n",
               "se calcularán las variables dummys")
  n <- ginput(message=msn,title = "ARBOLES DE DOS VARIABLES",
              icon = c("info"))
  n <- as.numeric(n)
  msn <- paste("Ingrese el número n de variables a interaccionar","\n",
               "Ejemplo: 5")
  cte <- ginput(message=msn,title = "VARIABLES A INTERACCIONAR",
                icon = c("info"))
  cte <- as.numeric(cte)
  PRBM_DUM_2VAR <- dummys_arbol2v (GB,BDD,indice=indicador_num[,1],n)
  View(PRBM_DUM_2VAR)

#### GRAFICO ARBOLES DE DECISION 1 VARIABLE CATEGORICA
```

```
#### GRAFICO DE SEDIMENTACIÓN PARA ELEGIR EL NUM DE VAR A CORTAR
graf_sedm(indicador_cat[,6],titulo="IND CAT")
print("MEDIANTE EL GRAF ELIJA EL NUMERO DE VAR CATEGORICAS A CORTAR")
msn <- paste("Ingrese el número n de variables categóricas para las cuales",
              "\n","se graficarán los arboles de decisión")
  n <- ginput(message=msn,title = "ARBOLES UNA VARIABLE CATEGORICA",
              icon = c("info"))
  n <- as.numeric(n)
  graf_arbol1v (GB,BDD,indice=indicador_cat[,1],n,tipo="categorica")
#### CALCULO PRBM, DUMMYS ARBOL 1 VARIABLE CATEGORICA
msn <- paste("Ingrese el número n de variables categóricas para las cuales",
             "\n","se calcularán las variables dummys")
  n <- ginput(message=msn,title = "ARBOLES UNA VARIABLE CATEGORICA",
              icon = c("info"))
  n <- as.numeric(n)
  PRBM_DUM_1VAR_CAT <- dummys_arbol1v (GB,BDD,indice=indicador_cat[,1],n)
  View(PRBM_DUM_1VAR_CAT)
#### GRAFICO ARBOLES DE DECISION 2 VARIABLES CATEGORICAS
msn <- paste("Ingrese el número n de variables para las cuales","\n",
              "se graficarán los arboles de decisión")
  n <- ginput(message=msn,title = "ARBOLES DOS VARIABLES CATEGORICAS",
              icon = c("info"))
  n <- as.numeric(n)
  graf_arbol2v (GB,BDD,indice=indicador_cat[,1],n,tipo="categorica")
  #### CALCULO PRBM, DUMMYS ARBOL 2 VARIABLES CATEGORICAS
  msn <- paste("Ingrese el número n de variables para las cuales","\n",
              "se calcularán las variables dummys")
  n <- ginput(message=msn,title = "ARBOLES DOS VARIABLES CATEGORICAS",
              icon = c("info"))
  n <- as.numeric(n)
  PRBM_DUM_2VAR_CAT <- dummys_arbol2v (GB,BDD,indice=indicador_cat[,1],n)
  View(PRBM_DUM_2VAR_CAT)
```

```
#### GRAFICO ARBOLES DE DECISION 3 VARIABLES CATEGORICAS
  msn <- paste("Ingrese el número n de variables para las cuales","\n",
               "se graficarán los arboles de decisión")
  n <- ginput(message=msn,title = "ARBOLES TRES VARIABLES CATEGORICAS",
              icon = c("info"))
  n <- as.numeric(n)
  graf_arbol3v (GB,BDD,indice=indicador_cat[,1],n,tipo="categorica")
#### CALCULO PRBM, DUMMYS ARBOL 3 VARIABLES CATEGORICAS
  msn <- paste("Ingrese el número n de variables para las cuales","\n",
               "se calcularán las variables dummys")
  n <- ginput(message=msn,title = "ARBOLES DE TRES VARIABLES CATEGORICAS",
              icon = c("info"))
  n <- as.numeric(n)
  PRBM_DUM_3VAR_CAT <- dummys_arbol3v (GB,BDD,indice=indicador_cat[,1],n)
  View(PRBM_DUM_3VAR_CAT)
#####################################################################
####                BASE CON PRBM Y DUMMYS                       ####
#####################################################################
bases <- c("PRBM_DUM_1VAR","PRBM_DUM_2VAR","PRBM_DUM_1VAR_CAT",
           "PRBM_DUM_2VAR_CAT","PRBM_DUM_3VAR_CAT")
b_d <- bases[which(bases %in% ls())]
if (length(b_d)>0){
  BDD_D_PRBM <- c(0)
  for (i in seq(1:length(b_d))){
    A <- get(b_d[i])
    BDD_D_PRBM <- data.frame(BDD_D_PRBM,A[,-1])
  }
}else{print("SE DEBE GENERAR AL MENOS UNA BASE DE VARIABLES DUMMYS")}
BDD_D_PRBM<-BDD_D_PRBM[,-1]
rm(list=c("PRBM_DUM_1VAR","PRBM_DUM_2VAR"))
rm(list=c("PRBM_DUM_1VAR_CAT","PRBM_DUM_2VAR_CAT","PRBM_DUM_3VAR_CAT"))
#### BDD_D_PRBM CON PRBM Y DUMMYS 1,2,3 VARIABLES NUM Y CAT
```

```
dim(BDD_D_PRBM)
View(BDD_D_PRBM)
write.table(file="BDD_PRBM_DUMMYS.csv",BDD_D_PRBM,dec=",",sep="\t")
#### A CONTINUACIÓN FILTRAREMOS BDD_D_PRBM
#####################################################################
####                FILTRADO DUMMYS Y PRBM                       ####
#####################################################################

#####################################################################
####           FUNCIONES FILTRADO DUMMYS Y PRBM                  ####
#####################################################################
####  FILTRO LA BASE BDD_D_PRBM
#### RAIZ DE LA MEDIA DE RESIDUOS AL CUADRADO PONDERADA (DIF TASA MALO)
RMR <- function(x,y,pesos=NULL){
  if(!is.null(pesos)){
    rmr <- sqrt(sum((pesos/sum(pesos))*((x-y)^2)))
  }else{
    rmr <- sqrt(sum(((x-y)^2))/length(x))
  }
  return(rmr)
}
#### FUNCION PARA CALCULO DE INDICADORES DUMMYS Y PRBM
ind_dummy <-function (GB,variable)
{
  vars <- data.frame(GB,variable)
  vars_m <- subset(vars,subset=vars[,1]==1)
  vars_b <- subset(vars,subset=vars[,1]==0)
  ## Estadistico KS
  ks <- dgof::ks.test(vars_m[,2],vars_b[,2],alternative="two.sided")
  ks <- round(as.numeric(ks$statistic),4)
  ## Coef de correlación de pearson
  corr <- round(abs(as.numeric(cor.test(vars[,1],vars[,2],method="pearson",
```

```
                conf.level = 0.95,na.action=getOption("na.action"))[4])),4)
  ## PRUEBA INDEPENDENCIA - COEF CONTINGENCIA
  chisq<-chisq.test(GB,variable)
  chisq<- as.numeric(chisq$statistic)
  cont <- round(sqrt(chisq/(chisq+length(variable))),4)
  ## ODDS VS % INICIAL DE MALO
  tc <- table(GB,variable)
  tot <- ifelse ((tc[1,] + tc[2,])==0,1,(tc[1,] + tc[2,]))
  odds <- tc[2,]/tot
  p <- as.vector (table(GB))
  porc_m <- rep(p[2]/sum(p),dim(tc)[2])
  rmr <- round(RMR(odds,porc_m,pesos=tot),4 )
  return(c(ks,corr,cont,rmr))
}
#####################################################################
####      LLAMADO FUNCIONES FILTRADO DUMMYS Y PRBM           ####
#####################################################################
#### ELIMINO DUMMYS CONSTANTES
const <- c(0)
for (i in seq(1:ncol(BDD_D_PRBM))){
  if(max(BDD_D_PRBM[,i])== min(BDD_D_PRBM[,i])){
    const <- c(const,i)}
}
const <- const[-1]
BDD_D_PRBM <- BDD_D_PRBM[,-(const)]
#### Se calculan los indicadores para Dummys y PRBM
indicador_dum <- numeric(6)
for(i in seq(1:ncol(BDD_D_PRBM))){
  ind <-c(i,names(BDD_D_PRBM)[i],ind_dummy(GB,BDD_D_PRBM[,i]))
  indicador_dum <- rbind(indicador_dum,ind)
}
#### INDICADORES PARA VARIABLES DUMMYS PROBABILIDADES DE MALO
```

```
indicador_dum <- rbind(indicador_dum[-1,])
indicador_dum <- data.frame (as.numeric (indicador_dum[,1]),
                             indicador_dum[,2],
                             as.numeric (indicador_dum[,3]),
                             as.numeric (indicador_dum[,4]),
                             as.numeric (indicador_dum[,5]),
                             as.numeric (indicador_dum[,6]))
IND <- 0.50*indicador_dum[,3]+0.10*indicador_dum[,4]+
  0.30*indicador_dum[,5]+0.10*indicador_dum[,6]
indicador_dum <- data.frame (indicador_dum,IND)
colnames(indicador_dum) <- c("COL_VARIABLE","VARIABLE","KS",
                             "ABS_CORR","CONT","RMRS","IND")
indicador_dum <- indicador_dum[order(indicador_dum[,7],
                                     decreasing=TRUE),]
View(indicador_dum)
## SE ACTUALIZA LA BASE PRIMERAS COLUMNAS VARS MAS PREDICTIVAS
BDD_D_PRBM <- BDD_D_PRBM[,indicador_dum[,1]]
#####################################################################
####       ELIMINACION DUMMYS Y PRBM CORR > 0.70             ####
#####################################################################
AUX <- cor(BDD_D_PRBM)
pos <- which(((abs(AUX)>=0.7) & (row(AUX) < col(AUX))),arr.ind=T)
col_elim <- numeric(nrow(pos))
for(i in seq(1:nrow(pos))){
  aux_col_elim <- c(pos[i,1],pos[i,2])
  if (!any(col_elim %in% aux_col_elim)){
    col_elim [i] <- pos[i,which.max(c(pos[i,1],
                                      pos[i,2]))]
  }
}
col_elim <- unique(col_elim[col_elim>0])
BDD_D_PRBM_FINAL <- BDD_D_PRBM[,-(col_elim)]
```

```
#names(BDD_D_PRBM_FINAL)
#### Se vuelve a calcular los indicadores para Dummys y PRBM finales
indicador_dum_f <- numeric(6)
for(i in seq(1:ncol(BDD_D_PRBM_FINAL))){
  if(max(BDD_D_PRBM_FINAL[,i])!= min(BDD_D_PRBM_FINAL[,i])){
    ind <-c(i,names(BDD_D_PRBM_FINAL)[i],ind_dummy(GB,BDD_D_PRBM_FINAL[,i]))
    indicador_dum_f <- rbind(indicador_dum_f,ind)
  }
  next
}
indicador_dum_f <- rbind(indicador_dum_f[-1,])
indicador_dum_f <- data.frame (as.numeric (indicador_dum_f[,1]),
                                    indicador_dum_f[,2],
                        as.numeric (indicador_dum_f[,3]),
                        as.numeric (indicador_dum_f[,4]),
                        as.numeric (indicador_dum_f[,5]),
                        as.numeric (indicador_dum_f[,6]))
IND <- 0.40*indicador_dum_f[,3]+0.15*indicador_dum_f[,4]+
  0.30*indicador_dum_f[,5]+0.15*indicador_dum_f[,6]
indicador_dum_f <- data.frame (indicador_dum_f,IND)
colnames(indicador_dum_f) <- c("COL_VARIABLE","VARIABLE","KS",
                                 "ABS_CORR","CONT","RMRS","IND")
indicador_dum_f <- indicador_dum_f[order(indicador_dum_f[,7],
                                           decreasing=TRUE),]
View(indicador_dum_f)
#### ELECCION DEL NUM DE DUMMYS
graf_sedm(indicador_dum_f[,7],titulo="IND TOTAL")
print("MEDIANTE EL GRAFICO ELIJA EL NUMERO DE VARIABLES DUMMYS")
write.table(file="INDICADOR_FINAL.csv",indicador_dum_f,dec=",",sep="\t")
#### VARIABLES REGRESION LOGISTICA
msn <- paste("Ingrese el número n de variables dummys y prbm",
                      "\n","a utilizar en regresión lgística")
```

```
n <- ginput(message=msn,title = "ARBOLES DE DOS VARIABLES",icon = c("info"))
n <- as.numeric(n)
####  BDD_DUMMY_PRBM_FINAL BASE CON PRBM Y DUMMYS MAS PREDICTIVAS
BDD_DUMMY_PRBM_FINAL <- BDD_D_PRBM_FINAL[,indicador_dum_f[(1:n),1]]
write.table(file="BDD_DUMMYS_PRBM_FINALES.csv",BDD_DUMMY_PRBM_FINAL,
                                                    dec=",",sep="\t")
rm(list=c("BDD_D_PRBM_FINAL"))
#####################################################################
####              FUNCIÓN REGRESIÓN LOGÍSTICA                   ####
#####################################################################
#base_regresion <- BR
Regresion_Logistica <- function (base_regresion)
{
  base_vect <- base_regresion
  #base_vect <- BR
  ### GB INDICADOR TIPO numeric
  BM<-base_vect[,1]
  ### GB debe ser factor para la regresión
  IBM<-as.factor(base_vect[,1])
  ### base_reg: BASE DE VAR PARA REGRESIÓN CON VAR DEP EN 1ra COL
  base_reg <- data.frame (IBM,base_vect[,-1])
  ### Partimos de la regresión inicial con todas las variables:
  ro <- glm (IBM~., data=base_reg[,-1], family = binomial("logit"))
  ### Seleccionamos mejor modelo en base al AIC:
  mejor_r <<- stepAIC (ro, direction = "backward")
  ### Numero regresores, excluyo el intercepto:
  num_reg <- as.numeric(mejor_r$rank) - 1
  ### Resumen del modelo
  summary(mejor_r)
  ### valor de AIC
  Akaike <- AIC(mejor_r)
#####################################################################
```

```
####         COEFICIENTES Y PRUEBAS ESTADISTICAS             ####
#####################################################################
####   COEFICIENTES
  ### Tabla completa de coef  (coef, error std, pvalor, etc)
  coeficientes <- as.data.frame(summary(mejor_r)$coefficients)
  ### Coeficientes exponenciales:
  coef_exp <- exp(coef(mejor_r))
  ### INTERVALOS DE CONFIANZA
  IC <- as.data.frame(confint(mejor_r,level = 0.95))
  ### IC Coeficientes exponenciales:
  IC_exp <- as.data.frame (exp(IC))
  VARIABLES_ECU <- data.frame(coeficientes,IC[,1],IC[,2],
                              coef_exp,IC_exp[,1],IC_exp[,2])
  colnames(VARIABLES_ECU) <- c("COEFICIENTE","ERROR STD.","Z-VALOR",
                               "SIGNIFICANCIA","LIM_INF","LIM_SUP",
                               "EXP (COEF)","LIM_INF_EXP","LIM_SUP_EXP")
  # a retornar
  #View(VARIABLES_ECU)
  ####    PRUEBAS ESTADISTICAS
  ## 2. PRUEBA DE WALD, (razón de verosimilitud)
  ## H0 : TODOS LOS COEF  = 0.
  ## H1 : TODOS LOS COEF !=0.
  ## SE RECHAZA H0 AL NIVEL 5%(ALPHA)  SI P-VALOR > 0.05.
  w <-regTermTest(mejor_r, mejor_r$formula, method="Wald")
  prueba_wald <- w$p
  prueba_wald
  # a retornar
  #prueba_wald
  ## 3. PRUEBA DE HOSMER-LESMESHOW
  ## H0: EL MODELO AJUSTA A LOS DATOS.
  ## H1: EL MODELO NO AJUSTA A LOS DATOS
  ## SE ACEPTA H0 AL NIVEL \ALPHA SI  p > 0.05
```

```
  ## 4. COEFICIENTE R CUADRADADO DE Nagelkerke
  R_cuad <- ORmultivariate(mejor_r)
  R_cuad_N <- R_cuad$Nagelkerke_R2
  # a retornar
  R_cuad_N
#####################################################################
####       MATRIZ DE CORRELACIÓN (MULTICOLINEALINAD)         ####
#####################################################################
####  MATRIZ CUADRADA DE ORDEN REGRESORES+1 (Incluido intercepto)
matriz_corr <- summary(mejor_r,correlation = TRUE)
 ####  Matriz de Correlaciones:
  CORRELACION <- as.data.frame(matriz_corr$correlation)
 ####  Matriz de varianzas y covarianzas:
  COV <- vcov(mejor_r)
 ####  PROBLEMA DE MULTICOLINEALIDAD.
 #### 1. VARIABLES CON CORRRELACIÓN   > 0.7
 ####  Elimino col de intercepto para analisis de MULTICOLINEALIDAD
  AUX <- CORRELACION[-1,-1]
  if(!is.null(dim(AUX))){
    P <- which((abs(AUX)>=0.7 & row(AUX) < col(AUX)),arr.ind=T)
    var_corr <- matrix (c(0),nrow=nrow (P), ncol=3)
    if (nrow(P)>0){
      print(paste("REGRESORES CORR > 0.70 EN:",
                  "Var_correlacionadas",sep=" "))
      for(i in seq(1:nrow(P)))
      {
        var_corr[i,1] <- row.names(AUX)[P[i,1]]
        var_corr[i,2] <- names(AUX)[P[i,2]]
        var_corr[i,3] <- AUX[P[i,1],P[i,2]]
      }
    }else{
      print("TODOS LOS REGRESORES SON INDEPENDIENTES")
```

```
    }
  } else{
    print("REGRESOR ÚNICO")
    var_corr <- "REGRESOR ÚNICO"
  }
#### 2. INDICE DE CONDICIONAMIENTO (IC)
#### IC = raiz(val_propio_max/val_propio_min) matriz correlacion(AUX)
v_p <- eigen(AUX, symmetric =TRUE, only.values = TRUE)
val_prop <- as.vector(v_p$values)
IC = sqrt(max(val_prop)/min(val_prop))
if(IC>15)
{print("MULTICOLINEALIDAD FUERTE")}
if(IC>=10 & IC<=15)
{print("MULTICOLINEALIDAD MODERADA")}
if(IC<10)
{print("NO HAY PROBLEMA DE MULTICOLINEALIDAD")}
### 3. FACTOR DE INFLACION DE LA VARIANZA (FIV) (para cada reg j)
### SI FIV_j > 10 Fuerte problema de MULTICOLINEALIDAD
FIV <- as.data.frame(diag(inv(AUX)))
colnames(FIV) <- c("FIV")
#View(FIV)
####  VERIFICACION
x <- cbind(row.names(FIV),FIV)
FIV_problem <- x[x[,2] > 10,]
if (nrow(FIV_problem) == 0){
print("NO HAY PROBLEMA DE MULTICOLINEALIDAD")
}else{
colnames(FIV_problem) <- c("VARIABLE","FIV PROBLEMA")
print("REGRESORES QUE CAUSAN MULTICOLINEALIDAD EN FIV_problem")
}
#View(FIV_problem)
#####################################################################
```

```
####           GRÀFICOS PARA CONFIRMAR SUPUESTOS             ####
#####################################################################
#### CONFIRMAMOS NORMALIDAD, HOMOCEDASTICIDAD
# E INDEPENDENCIA DE LOS ERRORES.
pdf(file = "GRAFICOS_REGRESION_LOGISTICA.pdf",width = 23, height = 7 )
#### Optional 1x2 graf/pág
par(mfrow=c(1,2))
plot(mejor_r)
#### Histograma residuos
residuos <- residuals(mejor_r)
hist(residuos,col="blue",border=FALSE,main="Histograma residuos",
   xlim=c(-5,5),ylim=c(0,1),probability=TRUE)
dev.off()
### Retornando varios objetos.
resul<-list(Variables_Ecuacion = VARIABLES_ECU,
          Matriz_Correlacion = CORRELACION,Matriz_Covarianza = COV,
          Var_Correlacionadas = var_corr,Factor_Infl_VAr = FIV_problem)
return(resul)
}
#####################################################################
####        LLAMADO FUNCIÓN REGRESIÓN LOGÍSTICA              ####
#####################################################################
## BR BASE PARA REGRESION faltan agregar var continuas
BR <- data.frame(GB,var_continuas1,BDD_DUMMY_PRBM_FINAL)
names(BR)
R <- Regresion_Logistica(BR)
####        RESULTADOS FUNCIÓN REGRESIÓN LOGÍSTICA           ####
R$Variables_Ecuacion
R$Matriz_Correlacion
R$Matriz_Covarianza
### En caso de no existir correlacion > 5% el data.frame
#Var_Correlacionadas resulta VACIO.
```

```
R$Var_Correlacionadas
### En caso de no existir MULTICOLINEALIDAD el data.frame
#FActor_Infl_VAr resulta VACIO.
R$FActor_Infl_VAr
####        RESULTADOS FUNCIÓN REGRESIÓN LOGÍSTICA           ####
View(R$Variables_Ecuacion)
View(R$Matriz_Correlacion)
View(R$Matriz_Covarianza)
write.table(R$Indicadores, file ="IND.csv" )
#####################################################################
####                 FUNCION RESULTADOS:                         ####
#####################################################################

#####################################################################
####               LIBRERIAS NECESARIAS                          ####
library(pROC)                                                     ####
require(PredictABEL)                                              ####
#####################################################################
resultados <- function (GB,mejor_r,corte=0.5){
#####################################################################
####        PROBABILIDADES PRONOSTICADAS Y SCORE                 ####
#####################################################################
#### Tabla de clasificación entre BM y la predicción P_BM:
  probabilidad <-predict(mejor_r,type="response")
  score <- round(1000*(1-probabilidad))
  P_GB <- as.numeric(ifelse((1-probabilidad)>corte,0,1))
  SCORE <- cbind(GB,P_GB,probabilidad,score)
  colnames(SCORE) <- c("GB","PREDICCION GB","PROBABILIDAD",
                       "SCORE")
  #View(SCORE)
  ### Tabla de clasificación entre GB y la predicción P_GB:
  T_C <- table (SCORE[,1],SCORE[,2],dnn = c("GB","P_GB"))
```

```
  T_C
  ### En % respecto al total de B/M iniciales (por fila:1)
  T_C_p <-prop.table(T_C,1)
  T_C_p
  T_C <- data.frame(T_C,round(T_C_p,2))
  colnames(T_C) <- c("GB","P_GB","FRECUENCIA","GB","P_GB",
                     "PORCENTAJE")
#####################################################################
####            KS, ROC, GINI DEL MODELO                         ####
#####################################################################
  SCORE_m <- subset(SCORE,subset=SCORE[,1]==1,4)
  SCORE_b <- subset(SCORE,subset=SCORE[,1]==0,4)
  ####  KS
  ks <- ks.test(SCORE_b,SCORE_m,alternative="two.sided")
  #pvalor_KS <- as.numeric(ks$p.value)
  ks <- as.numeric(ks$statistic)
  ####  ROC
  ROC <- roc(SCORE[,1], SCORE[,4],na.rm=TRUE,direction=c("auto"),
                                              auc=TRUE,plot=FALSE)
  AUC <- as.numeric(ROC$auc)
  ####  GINI
  GINI<- 2*AUC - 1
  INDICADORES <- c("KS","GINI","AUC")
  VALOR <- round(c(ks,GINI, AUC),4)
  INDICADORES <- data.frame(INDICADORES,VALOR)
#####################################################################
####             TABLAS DE PERFORMANCE                           ####
#####################################################################
  W <- data.frame(SCORE[,1],SCORE[,4])
  W <- W[order(W[,2],decreasing=FALSE),]
  ### vector de deciles
  d <- quantile(W[,2],probs=seq(0.1,1,0.1))
```

```
razon_bm <- c(0)
A <- matrix(ncol=11,nrow=10)
A[1,1] <- 1
A[10,2] <- 999
for(i in 1:(nrow(A)-1))
{
  A[i,2]<-  A[i+1,1]<- d[i]
}
TP <- data.frame(c(1:10),A)
for(i in 1:(nrow(TP)))
{
### clientes en cada decil
 TP[i,4] <- nrow (subset(W,subset = W[,2] > TP[i,2] & W[,2] <= TP[i,3]))
 TP[i,5] <- round ((TP[i,4]/nrow(W)),2)
### malos en cada decil
 TP[i,6] <- nrow (subset(W,
                 subset = W[,2] > TP[i,2] & W[,2] <= TP[i,3] & W[,1]==1))
 TP[i,7] <- round ((TP[i,6]/nrow(W)),2)
### buenos en cada decil
 TP[i,8] <- nrow (subset(W,
                subset = W[,2] > TP[i,2] & W[,2] <= TP[i,3] & W[,1]==0))
 TP[i,9] <- round ((TP[i,8]/nrow(W)),2)
### Razón exito / fracaso
 TP[i,10] <- round ((TP[i,6]/TP[i,4]),2)
 TP[i,11] <- round ((TP[i,8]/TP[i,4]),2)
 if (TP[i,8]>0)
 {
   r<-round((TP[i,6]/TP[i,8]),0)
   razon_bm[i] <- round((TP[i,6]/TP[i,8]),2)
   TP[i,12] <- paste(c(r,1),collapse=":")
 }else{
   r<-round(TP[i,6],0)
```

```
    razon_bm[i] <- TP[i,6]
    TP[i,12] <- paste(c(r,0),collapse=":")
  }
}
TP <- data.frame(TP[,1:5], cumsum(TP[,5]),TP[,6:7],cumsum(TP[,7]),
                        TP[,8:9],cumsum(TP[,9]),TP[,10:ncol(TP)])
colnames(TP) <- c("DECIL","DE","HASTA","N CLIE","% CLIE", "% ACUM CLIE",
                      "N EXIT","% EXIT","% ACUM EXIT","N FRAC","% FRAC",
                "% ACUM FRAC","% EXIT (DECIL)"  ,"% FRAC (DECIL)" ,"E/F")
 #View(TP)
#####################################################################
####             GRÁFICOS DISCRIMINACIÓN                         ####
#####################################################################
 pdf(file = "GRAFICOS_RESULTADOS.pdf",width = 23, height = 7 )
 par(mfrow=c(1,2))
 indiceb <- c(1:(nrow(SCORE)-sum(SCORE[,1])))
 indicem <- c(1:sum(SCORE[,1]))
 #### SCORE
 plot(SCORE_m,indicem,xlim=c(1,1000),col="red",xlab = "SCORE",
      ylab = "REGISTRO",main = "Score Buenos/Malos")
 points(SCORE_b,indiceb,col="green",ylim = c(0 ,sum(SCORE[,1])))
 #### TP
 #### spline: para interpolar puntos
 plot(spline(TP[,13],method = c("natural")),type="l",col="green",
                        ylim=c(0,1),xlab="DECIL",ylab="% E - F",
                                 main="DISTRIBUCIÓN POR DECIL")
 par(new=TRUE)
 plot(spline(TP[,14],method = c("natural")),type="l",col="red",
                       ylim=c(0,1),ylab="% E - F",xlab="DECIL")
 par(new=TRUE)
 plot((razon_bm-min(razon_bm))/max(razon_bm),type="h",xlim=c(1,10),
                col="blue",xlab="DECIL",ylim=c(0,1),ylab="% E - F")
```

```
  dev.off()
  ### Retornando varios objetos.
  h<-list(Score=SCORE,TP = TP,Tabla_Clasificacion = T_C,
          Indicadores=INDICADORES)
  return(h)
}
#####################################################################
####               LLAMADO FUNCIÓN RESULTADOS                    ####
#####################################################################
####  Cargamos el mejor modelo de regresión obtenido
mejor_regresion <- mejor_r
RESL <-  resultados(GB,mejor_r= mejor_regresion,corte=0.5)
View(RESL$Score)
View(RESL$TP)
View(RESL$Tabla_Clasificacion)
View(RESL$Indicadores)
```