

**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**



**FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS  
DEPARTAMENTO DE POSTGRADO**

**PROYECTO DE TITULACIÓN**

**PREVIO A LA OBTENCIÓN DEL TÍTULO DE:**

**“MAGÍSTER EN ESTADÍSTICA APLICADA”**

**TEMA:**

**MODELACIÓN ESTADÍSTICA DEL CUMPLIMIENTO DE PAGO  
PARA EL CRÉDITO PRODUCTIVO CON ACOMPAÑAMIENTO  
TÉCNICO COLOCADO POR UNA COOPERATIVA DE AHORROS**

**AUTOR:**

**PABLO ERNESTO VILLA GUERRERO**

**Guayaquil - Ecuador**

**2024**

## RESUMEN

El presente estudio tuvo como objetivo principal predecir el cumplimiento de pago de la cartera de créditos productivos vinculados a la Fundación CECCA y administrados por la Cooperativa de Ahorro y Crédito Jardín Azuayo, abordando una cartera considerable en riesgo. La investigación se basó en el análisis de 6262 transacciones provenientes de 18 oficinas distribuidas en cinco provincias de Ecuador, utilizando la metodología CRISP-DM. Se propuso un modelo predictivo basado en los días de morosidad en créditos liquidados, aplicando modelos de regresión como Poisson, binomial negativo y sus variantes infladas de ceros. El modelo de regresión binomial negativa inflada de ceros (ZINB) destaca al utilizar solo 13 variables significativas, evidenciando su eficacia con un AIC de 17552.843, un ECM de 4675.083, un RSME de 68.375, y un MAE de 40.476. Estas métricas indican que el ZINB logra una predicción precisa del cumplimiento de pago con un conjunto más eficiente y simplificado de predictores en comparación con otros modelos.

**Palabras claves:** morosidad, créditos, Poisson, binomial negativo, modelos inflados de ceros.

## **ABSTRACT**

The main objective of this study was to predict the payment compliance of the portfolio of productive loans linked to the CECCA Foundation and managed by the Jardín Azuayo Savings and Credit Cooperative, addressing a considerable portfolio at risk. The research was based on the analysis of 6262 transactions from 18 offices distributed across five provinces in Ecuador, using the CRISP-DM methodology. A predictive model based on the days of delinquency in settled loans was proposed, applying regression models such as Poisson, negative binomial, and their zero-inflated variants. The Zero-Inflated Negative Binomial (ZINB) regression model stands out by using only 13 significant variables, demonstrating its effectiveness with an AIC of 17552.843, an MCE of 4675.083, an RSME of 68.375, and an MAE of 40.476. These metrics indicate that the ZINB achieves precise payment compliance prediction with a more efficient and simplified set of predictors compared to other models.

Keywords: late payment, loans, Poisson, negative binomial, zero-inflated models.

## **DEDICATORIA**

A mis hijos, Kevin David Villa Quezada y Alan José Villa Quezada, quienes han sido mi mayor inspiración y motivación en este camino. Que este logro también sea un legado de perseverancia y determinación para ustedes. Los amo con todo mi corazón.

## AGRADECIMIENTO

Quisiera expresar mi más profundo agradecimiento a Dios, cuya guía y fortaleza me han acompañado en cada paso de este viaje académico. Sin su gracia y bendiciones, este logro no sería posible.

A mi madre, María Guerrero, por su apoyo incondicional. Su amor, sacrificio y palabras de aliento han sido el motor que me impulsó a seguir adelante en los momentos más desafiantes.

A mi esposa, Leidy Quezada García, mi compañera de vida, no tengo palabras suficientes para expresar mi gratitud. Su comprensión, paciencia y amor incondicional han sido mi soporte durante este proceso.

A la distinguida planta docente y administrativa del posgrado de la Facultad de Ciencias Naturales y Matemáticas de la ESPOL, así como a quienes estuvieron a cargo del Master en Estadística Aplicada, mi más sincero reconocimiento. Su dedicación y compromiso con la excelencia académica han sido fundamentales en mi formación profesional.

A mi tutor, Holger Cevallos Valdiviezo, PHD., agradezco enormemente su orientación experta, paciencia y apoyo incondicional a lo largo de este proceso. Sus consejos sabios y dirección académica fueron cruciales para el desarrollo y culminación de esta tesis.

Finalmente, a todos aquellos que de alguna manera contribuyeron a la realización de este trabajo, ¡muchas gracias! sus palabras de aliento, apoyo y motivación han sido invaluable para mí.

“Gracias por creer en mí”  
:)

## **DECLARACIÓN EXPRESA**

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación, me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría. El patrimonio intelectual del mismo, corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

---

Pablo Ernesto Villa Guerrero

# TRIBUNAL DE GRADUACIÓN

---

Galindo Villardón Purificación. Ph.D.  
PRESIDENTE

---

Cevallos Valdiviezo Holger Geovanny. Ph.D.  
TUTOR

---

Moreira Villegas Francisco Antonio. M.Sc.  
DOCENTE EVALUADOR

## ABREVIATURAS O SIGLAS

CECCA	Centro de Enseñanza y Educación del Campesinado del Azuay.
COAC	Cooperativa de Ahorro y Crédito.
UPA	Unidad de Producción Agropecuaria.
ONG	Organización No Gubernamental
CRISP-DM	Cross Industry Standard Process for Data Mining
IRR	Incidence Rate Ratio
RSME	Root Mean Squared Error
MCE	Error Cuadrático Medio
AIC	Criterio de información de Akaike
BIC	Criterio de información bayesiano
ZIP	Zero-Inflated Poisson
ZINB	Zero-Inflated Binomial Negativa
F	Estadística F de Fisher



# TABLA DE CONTENIDO

<b>CAPÍTULO 1</b> .....	<b>1</b>
<b>1. INTRODUCCIÓN</b> .....	<b>1</b>
1.1. Antecedentes .....	1
1.2. Descripción del problema .....	2
1.3. Objetivos .....	3
1.3.1. Objetivo general .....	3
1.3.2. Objetivos específicos.....	3
1.4. Alcance .....	3
<b>CAPÍTULO 2</b> .....	<b>4</b>
<b>2. MARCO TEÓRICO</b> .....	<b>4</b>
2.1. El crédito en el sector agrícola .....	4
2.2. Riesgo crediticio.....	4
2.3. Factores influyentes del riesgo crediticio.....	5
2.4. Modelación del riesgo crediticio.....	7
<b>3. MATERIALES Y MÉTODOS</b> .....	<b>9</b>
3.1. Tipo y diseño de investigación.....	9
3.2. Variable objetivo .....	9
3.2.1. Justificación de la variable.....	9
3.3. Metodología .....	9
3.3.1. Población objeto de estudio .....	10
3.3.2. Análisis del problema .....	10
3.3.3. Comprensión de los datos.....	10
3.3.3.1. Origen de los datos.....	10
3.3.3.2. Estructura de los datos .....	10
3.3.3.3. Verificación de la calidad de los datos.....	11
3.3.3.3.1. Valores perdidos.....	11
3.3.3.3.2. Valores duplicados .....	12
3.3.3.3.3. Consistencia de los datos y validación de valores.....	12
3.3.3.3.4. Valores atípicos u outliers .....	12
3.3.3.4. Exploración de los datos.....	12
3.3.4. Preparación de los datos.....	13
3.3.4.1. Selección de datos.....	13
3.3.4.2. Limpieza de datos.....	13
3.3.4.3. Construcción de nuevas variables.....	14
3.3.4.4. Formato de datos .....	16
3.3.5. Modelamiento .....	16
3.3.5.1. Selección de técnicas de modelado .....	16
3.3.5.2. Generación de un diseño de comprobación.....	16
3.3.5.3. Generación de los modelos .....	16
3.3.6. Evaluación .....	17
3.3.6.1. Evaluación de los resultados .....	17
<b>CAPÍTULO 4</b> .....	<b>18</b>

<b>4. RESULTADOS.....</b>	<b>18</b>
4.1. Base de datos .....	18
4.2. Análisis de la calidad de la base de datos.....	18
4.3. Descripción cada una de las variables asociadas con la morosidad y el cumplimiento de pago, mediante técnicas estadísticas descriptivas.....	36
4.3.1. Variables categóricas.....	36
4.3.2. Variables cuantitativas .....	38
4.3.3. Variable respuesta .....	41
4.3.3.1. Análisis descriptivo.....	41
4.4. Construcción modelos predictivos para el cumplimiento de pago, mediante técnicas estadísticas y de minería de datos. ....	44
4.4.1. Modelo de regresión Poisson .....	44
4.4.2. Modelo de regresión binomial negativa .....	47
4.4.3. Modelo de regresión Poisson inflada de ceros.....	49
4.4.4. Modelo de regresión binomial negativa inflada de ceros .....	54
4.5. Comparar el desempeño de cada uno de los modelos predictivos, utilizando una función de pérdida adecuada. ....	57
 <b>CAPÍTULO 5.....</b>	 <b>59</b>
<b>5. CONCLUSIONES Y RECOMENDACIONES.....</b>	<b>59</b>
5.1. Conclusiones .....	59
5.2. Recomendaciones .....	59
 <b>6. REFERENCIAS.....</b>	 <b>60</b>
 <b>7. APÉNDICES Y ANEXOS .....</b>	 <b>65</b>

## LISTADO DE FIGURAS

Fig. 4.1. Proporción y ubicación de valores perdidos (NA) en la base datos.....	20
Fig. 4.2. Diagrama de dispersión entre INGRESOS y GASTOS.....	24
Fig. 4.3. Boxplot de INGRESOS y GASTOS según EST_CIVIL .....	25
Fig. 4.4. Valores óptimos de lambda obtenidos por transformaciones Box-Cox. ....	27
Fig. 4.5. Boxplot de columnas financieras. ....	30
Fig. 4.6. Boxplot de columnas financieras con transformación Box-Cox.....	30
Fig. 4.7. Boxplot de columnas financieras después eliminación de outliers. ....	31
Fig. 4.8. Numero de creditos por territorio. ....	32
Fig. 4.9. Distribución de los niveles de INSTRUCCION.....	33
Fig. 4.10. Histograma y curva de densidad para N_CARGAS.....	33
Fig. 4.11. Histograma y curva de densidad para PLAZO.....	34
Fig. 4.12. Histograma y curva de densidad para TASA_INTERES.....	34
Fig. 4.13. Número de creditos por FREC_INGRESOS.....	35
Fig. 4.14. Número de creditos por TIPO.....	35
Fig. 4.15. Número de creditos por EST_CREDITO.....	36
Fig. 4.16. Distribución de variables categóricas.....	37
Fig. 4.17. Histograma de columnas financieras con transformación Box-Cox.....	39
Fig. 4.18. Visualización de la variable respuesta DMORA.....	42
Fig. 4.19. Familias de distribución de la morosidad (DMORA).....	43

## LISTADO DE TABLAS

Tabla 2.1 Calificación de la cartera de crédito para Cooperativas de Ahorro y Crédito y Asociaciones Mutualistas de Ahorro y Crédito para la Vivienda. ....	5
Tabla 2.2. Características del conjunto de datos inicial .....	10
Tabla 2.3. Codificación y tipos de variables de la base de datos .....	11
Tabla 2.4. Variables consideradas para modelamiento del riesgo crediticio .....	15
Tabla 4.1. Proporción de valores perdidos por variable .....	19
Tabla 4.2. Proporciones y numero de valores duplicados por columna. ....	21
Tabla 4.3. Valores únicos para columnas categóricas.....	22
Tabla 4.4. Valores únicos para columnas numéricas.....	23
Tabla 4.5. Resumen estadístico de variables cuantitativas.....	26
Tabla 4.6. Valores atípicos y normales según transformaciones Box-Cox.....	29
Tabla 4.7. Estadística descriptiva para variables cuantitativas. ....	40
Tabla 4.8. Estadísticas descriptiva de la variable días de morosidad.....	41
Tabla 4.9. Resumen del modelo de regresión Poisson.....	44
Tabla 4.10. Parámetros del modelo de regresión Poisson.....	47
Tabla 4.11. Resumen del modelo de regresión binomial negativa.....	48
Tabla 4.12. Parámetros del modelo de regresión binomial negativa.....	49
Tabla 4.13. Resumen del modelo de regresión Poisson inflada de ceros. ....	50
Tabla 4.14. Parámetros del modelo de regresión Poisson inflado con ceros.....	53
Tabla 4.15. Resumen del modelo de regresión binomial negativa inflada de ceros.....	54
Tabla 4.16. Parámetros del modelo de regresión binomial negativa inflado con ceros. ...	56
Tabla 4.17. Métricas de comparación de modelos de regresión POI, BN, ZIP y ZINB. ....	58
Tabla 5.1 Análisis descriptivo de las variables independientes cuantitativas .....	66

## LISTADO DE ANEXOS

Anexo A. Características y detalles de la base de datos .....	65
Anexo B. Características y validación del conjunto de datos .....	67
Anexo D. Estructura de base de datos .....	70
Anexo E. Partición de base de datos y construcción de modelos de regresión .....	71

# Capítulo 1

## 1. INTRODUCCIÓN

### 1.1. Antecedentes

En 1979, se crea el Centro de Educación y Capacitación del Campesinado del Azuay (CECCA) a consecuencia de los eventos de la reforma agraria que afrontaba Ecuador [1], [2]. El CECCA se constituyó motivado a generar apoyo en los procesos de capacitación a los poblados que nacieron a raíz de la toma de la Hacienda “Tenguel”, mediante diversos proyectos propuestos por la Unión Regional de Organizaciones Campesinas del Litoral [2].

En febrero de 1996, a raíz de la crisis provocada por el desastre de La Josefina en el cantón Paute; por gestión de la Fundación CECCA se viabilizó el proyecto de instituir a la Cooperativa de Ahorro y Crédito Jardín Azuayo. La Cooperativa al inicio de sus actividades identificó que en los cantones de Paute y Gualaceo existía un sistema de financiamiento excluyente del crédito a pequeños productores agropecuarios; de manera que, generaron finanzas adaptadas a las características del campo, con el fin de permitirles adquirir tecnologías apropiadas para reducir los costos y fomentar el desarrollo económico productivo [3], [4].

Para el 2002, la Fundación CECCA crea el programa denominado “Créditos Productivos con Acompañamiento Técnico” en colaboración con la Cooperativa Jardín Azuayo y otras ONG’s, con la finalidad de fortalecer el acceso al crédito para el pequeño productor agropecuario excluido del sistema financiero tradicional, con cobertura en ciertos cantones de las provincias de Azuay, Cañar, Loja, Guayas y El Oro [2], [3], [4], [5].

En la actualidad, la sostenibilidad del Programa de Créditos Productivos se ha mantenido gracias a la gestión eficiente de préstamos durante las últimas dos décadas. Sin embargo, en marzo de 2022, se registró un riesgo crediticio del 5.9%. Para garantizar la continuación de la participación de los agricultores en los procesos de inclusión financiera a través del Programa, la Fundación CECCA debe mantener un índice de morosidad por debajo del registrado por la entidad que administra sus recursos financieros, la Cooperativa de Ahorro y Crédito Jardín Azuayo, que asciende al 3.15% [6]. Esto implica la necesidad de implementar estrategias efectivas para reducir el riesgo crediticio y asegurar la estabilidad financiera del programa en el futuro.

Predecir oportunamente aquellos socios con una mala conducta en los créditos es una necesidad que ha ido en aumento; es por esto que, las reguladoras financieras buscan mejorar frecuentemente el modelo predictivo de incumplimiento que han adoptado, ya que saben que el riesgo crediticio es una pérdida importante dentro de la administración financiera [7].

## **1.2. Descripción del problema**

Las cooperativas de ahorro y crédito son parte importante en el desarrollo económico del país, más aún para el sector de la agricultura, ya que es una de las actividades económicas más importantes de Ecuador, ya que ha permitido el desarrollo del sistema económico de las industrias del sector [8], [9]. Según el Banco Central, en el 2020 el comercio exterior agropecuario representó el 50,41% de las exportaciones no petroleras ecuatorianas; es decir, la entrada de USD 7.549 millones al país [10].

Debido al rol que cumplen las empresas agropecuarias dentro del ambiente económico, estas deben ser más competitivas, lo que provoca que estén sujetas a cambios o mejoras tecnológicas o sociales; por lo que, buscan el financiamiento que en ocasiones suele ser una limitación, puesto que muchos productores agrícolas no logran cumplir con los estándares de garantías y los múltiples requerimientos que solicitan las entidades financieras para el acceso a créditos. El sector agropecuario es considerado como de alto riesgo para el financiamiento debido al grado de informalidad de los predios agrícolas, la variabilidad de los precios o la dependencia a factores climáticos y de mercadeo. Debido a esto, algunas entidades financieras les ofrecen a los agricultores un producto crediticio basado en sus recursos y que respondan adecuadamente a las necesidades y particularidades que requiera; de manera que, les permita adquirir tecnología, maquinarias, equipos e innovaciones en sus procesos productivos [11], [12], [13], [14].

El riesgo comienza en el momento mismo de aprobar el crédito y desembolsar el dinero, las circunstancias que pueden afectar el pago oportuno son múltiples; por lo tanto, es necesario para una institución financiera analizar y evaluar el riesgo que conlleva la otorgación de créditos, ya que si no se realiza una gestión óptima al momento de seleccionar a los socios beneficiarios de recursos financiados, la institución incurriría en un incremento en el índice de morosidad [15], [16], [17], [18].

Los objetivos de las Cooperativas de Ahorro y Crédito dependen de la gestión eficiente de sus recursos. Por ende, se considera en lo posible identificar las variables de los factores agroproductivos vinculados a la Unidad de Producción Agropecuaria (UPA) y los factores económicos y sociales de los socios, que

guarden relación con la morosidad de tal forma que podamos modelar su cumplimiento de pago, para generar una intervención oportuna y minimizar la probabilidad del incumplimiento de los beneficiarios del crédito [19].

### **1.3. Objetivos**

#### **1.3.1. Objetivo general**

Predecir el cumplimiento de pago basado en los días de mora de los créditos productivos otorgados por la Cooperativa de Ahorro y Crédito Jardín Azuayo, mediante el uso de diferentes técnicas estadísticas y de minería de datos, que permita una mejor estrategia en el proceso de otorgación de créditos.

#### **1.3.2. Objetivos específicos**

- Describir cada una de las variables asociadas con la morosidad y el cumplimiento de pago, mediante técnicas estadísticas descriptivas.
- Construir modelos predictivos para el cumplimiento de pago, mediante técnicas estadísticas y de minería de datos.
- Comparar el desempeño de cada uno de los modelos predictivos, utilizando una función de pérdida adecuada.
- Generar predicciones de la morosidad con el fin de anticipar el incumplimiento de pago.

### **1.4. Alcance**

En este estudio se analizó las características sociales y financieras del productor, las condiciones del predio agrícola y los indicadores financieros provenientes de las transacciones de créditos otorgados desde enero del 2002 hasta noviembre del 2023; de aquellos productores provenientes de los cantones en los que la Fundación CECCA presta atención a nivel nacional como Naranjal, La Troncal, Azogues, Oña, Ponce Enríquez, Pasaje, Cañar, Paute, Santa Rosa, Sígsig, Machala, Palmas, Sevilla de Oro, Saraguro, Gualaceo y Chordeleg.



# Capítulo 2

## 2. MARCO TEÓRICO

### 2.1. El crédito en el sector agrícola

El acceso al crédito es un elemento crucial para el crecimiento y la productividad de los pequeños agricultores, desempeñando un papel fundamental en diversos aspectos del desarrollo económico. No sólo facilita la adopción de nuevas tecnologías, sino también contribuye al aumento de la productividad y la mejora general del nivel de vida del agricultor [20].

Por esta razón, en países en desarrollo como Nigeria, se presenta una baja productividad, principalmente debido al escaso empleo de tecnologías agrícolas modernas, tales como fertilizantes, semillas mejoradas, agroquímicos y mecanización. Esta situación se atribuye al acceso deficiente al crédito que enfrentan los agricultores en la región [21]. A diferencia de lugares como la provincia de Heilongjiang en China, donde convencidos de la necesidad de los créditos agrícolas, lograron un crecimiento en la producción agrícola del 0.12%, tan solo en aumentar los créditos agrarios en un 1%, ya que consideran a las finanzas como el centro de la economía agrícola [22].

### 2.2. Riesgo crediticio

El riesgo de crédito o riesgo de incumplimiento de pago implica la incapacidad o falta de voluntad de un cliente o contraparte para cumplir con los compromisos en relación con préstamos y otras transacciones financieras [18], [23]. En el caso de los agricultores, el incumplimiento de pago puede originarse debido a diversas situaciones, ya que la producción agrícola está expuesta a una serie de riesgos, entre los que se incluyen los riesgos climáticos, los riesgos de mercado y los riesgos de producción. Estos riesgos pueden tener un impacto directo en la rentabilidad de los productores [24], [25].

Es por esto, que las instituciones financieras suelen atribuir que el incumplimiento de pago se debe al hecho de prestar a clientes más riesgosos, en lugar de ser consecuencia de una ineficiencia en el otorgamiento del crédito [26]. Cuando se presenta el incumplimiento respecto de las condiciones iniciales establecidas en el crédito, ya sea en términos de la fecha de pago o el monto de la cuota, el crédito entra en mora [27].

La morosidad es un indicador clave del riesgo crediticio, ya que puede llevar a pérdidas financieras significativas para la institución prestamista [28]. Dado que la morosidad está estrechamente relacionada con la rentabilidad de la institución financiera, su medición se vuelve esencial para comprender su impacto en la estabilidad financiera y respaldar la toma de decisiones informadas [29].

La morosidad se puede medir a través del porcentaje de clientes con deudas vencidas y gestionadas con 30 días de antelación o menos [30]. También, se puede calcular a través de una tasa o ratio, dividiendo el valor de los créditos en riesgos entre el valor total de la cartera de créditos [31]. Además, existen otras métricas, como el ratio de mora, la tasa de cobertura y el coste del riesgo [32]. La medición de la morosidad se realiza comúnmente en días o en términos de tasa porcentual [33].

Según los días de mora se puede identificar cinco tipos de riesgo: normal, potencial, deficiente, dudoso recaudo y pérdida [34]. Hasta el 31 de diciembre de 2022, las entidades financieras calificarán la cartera de crédito conforme a la Tabla 2.1.

**Tabla 2.1**  
**Calificación de la cartera de crédito para Cooperativas de Ahorro y Crédito y Asociaciones Mutualistas de Ahorro y Crédito para la Vivienda.**

Nivel de riesgo	Categoría	Productivo	Consumo, microcrédito, educativo	Inmobiliario, vivienda de interés social y publico
		Días de morosidad		
Riesgo normal	A1	0-5	0-5	0-5
	A2	6-30	6-30	6-45
	A3	31-60	31-60	46-90
Riesgo potencial	B1	61-75	61-75	91-150
	B2	76-90	76-90	151-210
Riesgo deficiente	C1	91-120	91-120	211-270
	C2	121-180	121-150	271-360
Dudoso recaudo	D	181-360	151-180	361-450
Perdida	E	Mayor a 360	Mayor a 180	Mayor a 450

Nota: Valores tomados de la Reforma a las Normas para la Gestión del Riesgo de Crédito y para la Constitución de Provisiones en las Cooperativas de Ahorro y Crédito y Asociaciones Mutualistas de Ahorro y Crédito para la Vivienda [35].

### 2.3. Factores influyentes del riesgo crediticio

Es importante mantener un índice de morosidad bajo control, ya que esto impulsa el crecimiento continuo del sistema financiero [36]. Cuando esto no se cumple, surge el riesgo crediticio, componente crítico en la gestión financiera. Este riesgo

se ve afectado por diversos factores, tanto internos como externos, y varía entre instituciones debido a las diferencias en sus carteras de crédito [37]. Estos factores, que inciden directamente en la capacidad de los prestatarios para cumplir con sus compromisos financieros, evidencian la necesidad de una gestión cuidadosa de la cartera para mantener la estabilidad financiera [38].

En un estudio destinado a evaluar el impacto de los riesgos de no pago entre agricultores de soja en Togo, se identificaron diversos factores que inciden significativamente en la percepción del riesgo crediticio. En este sentido, la edad de los agricultores, la afiliación a una organización, el tipo de cultivo, la forma de comercialización de la soja, junto con el género del agricultor y el monto del crédito, surgen como indicadores clave en las restricciones crediticias [39]. Resultados similares se vieron en agricultores de maíz en Ghana, , donde también influyeron aspectos como el género, la pertenencia a una organización social y la voluntad de pago por parte del agricultor [40].

Möllmann et al. [36] sostienen que la vulnerabilidad de los agricultores a los fenómenos meteorológicos desfavorables, que probablemente aumentarán en frecuencia y magnitud debido al cambio climático, es un impedimento importante para que los prestamistas, ya que puede generar impagos durante condiciones climáticas severas.

En [41] se investigó el efecto económico de la pandemia de COVID-19 sobre el riesgo crediticio soberano utilizando datos de cuarenta países desarrollados y en desarrollo. Los resultados de este estudio revelaron un aumento significativo en el riesgo crediticio soberano después del brote de la pandemia, debido a que genera serias preocupaciones sobre la contracción del consumo agregado, además de la capacidad fiscal y la volatilidad de las exportaciones. Además, determinaron que el riesgo crediticio soberano tiende a aumentar aún más en aquellos países que experimentan una disminución en las ventas minoristas durante la pandemia o que presentan una tasa de ahorro agregado más baja antes del brote.

Shahab et al. [42] analizaron los efectos de factores socioeconómicos en el riesgo de incumplimiento crediticio de los agricultores. Los resultados demostraron que la educación, la experiencia, el tipo de agricultor (ya sea de subsistencia o empresarial), el tamaño de la propiedad, los ingresos no agrícolas y los riesgos de inundaciones tienen un impacto significativo en el riesgo crediticio.

El gobierno de Pakistán evaluó los factores que inciden en el riesgo crediticio de los agricultores en sus programas de crédito y determinó que el factor educación tiene un impacto significativo en el riesgo crediticio. Además, observaron que los ingresos transitorios también desempeñan un papel crucial en el comportamiento crediticio, al igual que los activos iniciales del agricultor [43].

Factores como el área geográfica, la línea de negocio, el sector económico y el tipo de cartera, son elementos que, desempeñan un papel esencial al permitir a las instituciones financieras no solo identificar, sino también medir y administrar de manera efectiva el riesgo crediticio [44].

## **2.4. Modelación del riesgo crediticio**

La evaluación del riesgo crediticio ha motivado a varios investigadores a desarrollar modelos de clasificación para predecir el riesgo de los prestatarios en función de diversas variables tanto financieras como no financieras [45], [46]. Estos modelos emplean técnicas estadísticas para estimar la probabilidad de incumplimiento de pago por parte de los prestatarios [46].

En este sentido, destacan los estudios de Wagdi y Tarek [47], quienes utilizaron redes neuronales artificiales (ANN por sus siglas en inglés) para clasificar el riesgo de los prestatarios con bancos egipcios, utilizando datos financieros de estos bancos y de la economía local, logrando explicar el 23.31% de variación en el crédito del deudor de los prestatarios. Así mismo, en [48] mediante el uso de una red neuronal se logró una precisión de clasificación del 79.59% del riesgo crediticio en cooperativas de ahorro y crédito del Ecuador, en base a índices financieros de los clientes.

Otros estudios han evaluado otros métodos para clasificar el riesgo al momento de conceder préstamos. Por ejemplo, en [49], se utilizan modelos de análisis discriminante y de regresión logística para clasificar a clientes en base a su riesgo crediticio, logrando una precisión de clasificación del 88.9% para el análisis discriminante y del 94.4% para la regresión logística. Los autores en [50] destacan también los modelos de regresión logística para la clasificación del riesgo de clientes.

Otra técnica utilizada en la literatura para la clasificación del riesgo crediticio es la de bosques aleatorios. Por ejemplo, el trabajo de Paula et al. [19], quienes, con el fin de aumentar la efectividad del proceso de concesión de préstamos en cooperativas de crédito brasileñas, modelaron la calificación crediticia usando bosques aleatorios y regresión logística. En este estudio, la técnica de bosques aleatorios muestra un mejor rendimiento de predicción de la calificación crediticia que el modelo de regresión logística. Los autores además indican que utilizar los resultados de ambas técnicas de forma conjunta puede contribuir a una toma de decisiones más eficaz.

Los estudios anteriormente mencionados clasifican la solvencia crediticia de los prestatarios según la probabilidad estimada de incumplimiento. Sin embargo, es

posible también realizar la clasificación crediticia de los clientes en base a la predicción del número de días de morosidad, tal como lo mencionamos anteriormente [51]. Modelar los días de mora también puede ser de utilidad para identificar patrones y tendencias en el comportamiento de pago, con el fin de ayudar a las instituciones financieras a tomar decisiones más informadas sobre la gestión de riesgos y la asignación de recursos [52].

Al usar el número de días de mora como variable de respuesta en el modelo predicción, puede ser común la presencia en los datos de entrenamiento de muchos clientes con cero días de mora, es decir que tienen ausencia de retraso en el pago y por ende realizan el pago de forma puntual [53]. La ocurrencia de cero días de mora en instituciones financieras tiende a ser significativamente elevada en comparación con otros intervalos de mora, especialmente en aquellas entidades financieras que presentan bajos índices de morosidad [54].

En estos casos, es imprescindible tomar en cuenta el exceso de ceros en la variable de respuesta al momento de construir el modelo de predicción. Para ello, pueden emplearse modelos inflados de ceros [55]. Estos procedimientos combinan un modelo que considera la presencia de exceso de ceros en la variable de días de morosidad con otro modelo que considera el conteo per se del número de días de morosidad [56].

Para modelar los días de morosidad, es posible utilizar modelos para datos de conteo tales como la regresión Poisson o la regresión binomial negativa. La regresión Poisson modela el conteo de los días de morosidad, asumiendo que la media y la varianza de los días de morosidad son iguales. Esto generalmente es un supuesto muy restrictivo. Por su parte, la regresión binomial negativa puede ser usada cuando este supuesto no es justificable en los datos, ya que modela la sobredispersión de la variable de respuesta, que para este estudio es los días de morosidad. Tanto la regresión Poisson como la regresión binomial negativa poseen variantes para el modelamiento de la inflación de ceros en la variable de respuesta [57].

La importancia de crear modelos de morosidad radica en la capacidad de identificar los factores que afectan a la tasa de morosidad en instituciones financieras. Amat, Manini, y Renart [58] recomiendan que también se debe considerar el factor humano en el proceso de otorgamiento de crédito y los modelos estadísticos deben ser usados como una herramienta para guiar la toma de decisiones, complementaria al factor humano y la experiencia.

# Capítulo 3

## 3. MATERIALES Y MÉTODOS

### 3.1. Tipo y diseño de investigación

Para llevar a cabo este estudio, se adoptó una metodología de investigación mixta que permitió el análisis tanto de variables cuantitativas como cualitativas. Esta elección metodológica se basa en la necesidad de construir un modelo de riesgo crediticio usando los días de mora como variable de respuesta y las diversas características sociales, financieras y agrícolas presentes en las solicitudes de créditos como variables predictoras.

### 3.2. Variable objetivo

La métrica principal considerada para este estudio fue la cantidad de días de retraso en los pagos de los créditos (días de morosidad). Esta variable se seleccionó por su importancia crítica en la evaluación del desempeño de las entidades financieras, ofreciendo una medida clara del grado de incumplimiento y retraso en los pagos por parte de los beneficiarios de los créditos [59].

#### 3.2.1. Justificación de la variable

La elección de analizar los días de morosidad en cada solicitud de crédito cancelada se fundamenta en las directrices establecidas por la Junta de Política y Regulación Monetaria y Financiera (JRPF) en Ecuador. Según estas directrices, los créditos que presentan morosidad a partir de 61 días son considerados un riesgo potencial [34]. Esta investigación propone obtener predicciones de la cantidad de días de mora con el que el cliente culmina el pago del crédito. Esta es una contribución crucial para la Fundación CECCA, ya que les puede permitir gestionar de manera más efectiva el riesgo crediticio y con el fin de mantener el porcentaje de cartera en riesgo por debajo del total de la Cooperativa Jardín Azuayo.

### 3.3. Metodología

Para el análisis de los datos del estudio se utilizó la regresión Poisson y la regresión binomial, junto con sus variantes para modelar la inflación de ceros, las cuales fueron implementadas en el lenguaje de programación R.

### 3.3.1. Población objeto de estudio

Se analizaron 6262 transacciones crediticias provenientes de 18 oficinas de la Cooperativa de Ahorro y Crédito Jardín Azuayo, distribuidos en cinco territorios, ubicados en las provincias de El Oro, Guayas, Cañar, Azuay y Loja, tal como se detalla en el Anexo A:

Tabla A1.

### 3.3.2. Análisis del problema

Se ha identificado una situación preocupante en la Cooperativa Jardín Azuayo con el producto “Crédito Productivo con Acompañamiento Técnico”, el cual es promocionado y administrado por la Fundación CECCA. El problema principal radica en que la cartera en riesgo supera el valor total de la Cooperativa. Como estrategia para abordar esta situación, se ha propuesto el desarrollo de varios modelos de cumplimiento de pago usando como variable de respuesta los días de morosidad en créditos previamente liquidados por los socios.

### 3.3.3. Comprensión de los datos

La base de datos para este estudio fue suministrada por la Cooperativa Jardín Azuayo a través de la Fundación CECCA.

#### 3.3.3.1. Origen de los datos

Los datos utilizados en este estudio provienen del Reporte de “Fondos Administrados y Crédito” emitido por la Cooperativa de Ahorro y Crédito Jardín Azuayo, con un corte hasta marzo de 2022. Los datos fueron recibidos en un archivo de Microsoft Excel (.xlsx).

#### 3.3.3.2. Estructura de los datos

La base de datos de origen consta de 6262 registros de créditos que han sido otorgados a los socios en el pasado. Cada registro está vinculado con 31 columnas, las cuales contienen información detallada sobre transacciones crediticias de cada socio, como se detalla en la Tabla 2.2.

**Tabla 2.2.**  
**Características del conjunto de datos inicial**

Filas	Columnas	Columnas		Valores totales perdidos	Observaciones totales	Memoria
		Discretas	Continuas			
6262	31	15	16	10962	194122	2.56 MB

Nota: Características de la base de datos del Reporte de Fondos Administrados y Crédito de la Cooperativa de Ahorro y Crédito Jardín Azuayo.

La composición inicial de la base de datos incluía 15 columnas de tipo carácter y 16 columnas de tipo numérico, que representaban el 48.39% y 51.61% del total

de columnas, respectivamente. Para facilitar el análisis, todas las columnas del conjunto de datos inicial fueron codificadas, como se detalla en la Tabla 2.3.

**Tabla 2.3.**  
**Codificación y tipos de variables de la base de datos**

n	Id. Origen	Id. Modificado	Característica	Tipo
1	Socio	COD_SOCIO	Código del socio	Numérico
2	EDAD	EDAD	Edad	Numérico
3	GENERO	GENERO	Genero	Carácter
4	INSTRUCCIÓN	INSTRUCCION	Nivel de instrucción	Carácter
5	ESTADO.CIVIL	EST_CIVIL	Estado civil	Carácter
6	CARGAS	N_CARGAS	Numero de cargas	Numérico
7	ACTIVOS	ACTIVOS	Activos	Numérico
8	PASIVOS	PASIVOS	Pasivos	Numérico
9	INGRESOS	INGRESOS	Ingresos	Numérico
10	GASTOS	GASTOS	Gastos	Numérico
11	Nombre	NOMBRE_SOCIO	Nombre	Carácter
12	Sucursal	SUCURSAL	Sucursal	Carácter
13	Territorio	TERRITORIO	Territorio	Carácter
14	Credito	COD_CREDITO	Código del credito	Numérico
15	Est.Credito	EST_CREDITO	Estado del credito	Carácter
16	F.Concesion	FECHA_CONCESION	Fecha concesión del credito	Carácter
17	F.Vencimiento	FECHA_EVNCIMIENTO	Fecha vencimiento del credito	Carácter
18	Plazo	PLAZO	Plazo	Numérico
19	Cantidad	MONTO	Monto	Numérico
20	Tasa.Int	TASA_INTERES	Tasa de interés	Numérico
21	Desrtino	DESTINO	Destino	Carácter
22	Saldo_Total	SALDO_TOTAL	Saldo total del credito	Numérico
23	Saldo_Vencido	SALDO_VENCIDO	Saldo vencido del credito	Numérico
24	Calif	CALIFICACION	Calificación de morosidad	Carácter
25	Diasmora	DIAS_MORA_AC	Días de mora acumulado	Numérico
26	Pignoracion	PIGNORACION	Pignoración	Numérico
27	Int..Pag	INTERES_PAGADO	Cantidad de interés pagados	Numérico
28	Observacion	OBSERVACION	Observación	Carácter
29	Tipo	TIPO	Tipo diferimiento de credito	Carácter
30	Actividad	ACTIVIDAD_AGRO	Actividad agropecuaria	Carácter
31	Grupo.agricola	GRUPO_AGRO	Grupo agropecuario	Carácter

Nota: Identificar (Id.) de columnas disponibles en la base de datos original. n = posición de la columna en la base de datos.

### 3.3.3.3. Verificación de la calidad de los datos

Se llevó a cabo una evaluación preliminar de la calidad de los datos, en la cual se identificó posibles valores atípicos, valores faltantes y anomalías.

#### 3.3.3.3.1. Valores perdidos

La evaluación de valores perdidos en la base de datos se llevó a cabo mediante la función `ezPrecis()` de la librería "ez" en R. Las columnas sin valores faltantes no necesitaron intervención adicional. En el caso de las columnas con menos del 10% de datos faltantes, se optó por aplicar técnicas de imputación mediante métodos de regresión respectivos al tipo de variable. Sin embargo, para aquellas



columnas con más del 50% de datos faltantes, se consideró la exclusión de dichas variables del análisis.

#### **3.3.3.3.2. Valores duplicados**

Para la detección de valores duplicados, se empleó el paquete `dplyr` en R para analizar la presencia de valores duplicados en el conjunto de datos. Se utilizó la función `summarise_all()` junto con `duplicated()` en todas las columnas de la base de datos. Esta metodología identificó duplicados en cada columna, generando valores lógicos que indicaban la repetición de cada valor.

En el análisis de valores duplicados, se focalizó en tres columnas: "COD\_SOCIO", "NOMBRE\_SOCIO" y "COD\_CREDITO", fundamentales en la gestión de los créditos otorgado. Se esperaba que las dos primeras tuvieran una cantidad similar o idéntica de valores duplicados, mientras que la última, como indicadora de las solicitudes de créditos, no debía contener duplicados. La identificación de un duplicado en "COD\_CREDITO" condujo a la eliminación de la fila correspondiente.

#### **3.3.3.3.3. Consistencia de los datos y validación de valores**

Considerando que la mayoría de los datos se ingresaban manualmente, se llevaron a cabo resúmenes de dimensiones mediante las funciones de las librerías `dplyr` y `tidyr` para cada columna del conjunto de datos. Esto tenía como fin detectar errores tipográficos, de medición o de escalas, así como verificar que las fechas estuvieran en el formato adecuado y que los datos numéricos fueran coherentes.

#### **3.3.3.3.4. Valores atípicos u outliers**

Para analizar los valores outliers se realizó una visualización de la dispersión de los datos mediante el uso de `geom_boxplot()` de la librería `ggplot`. Esta visualización permitió identificar de manera inicial la presencia de posibles valores atípicos en las diferentes columnas de interés. Posteriormente, se establecieron umbrales basados en el Rango Intercuartílico (IQR) para detectar valores atípicos que se ubicaran fuera de los límites establecidos. La estrategia adoptada para abordar los valores atípicos fue la eliminación de las observaciones que superaban dichos umbrales. Esta decisión se basó en el análisis específico de cada variable afectada, la naturaleza propia de los datos y su importancia relativa para los objetivos de la investigación.

#### **3.3.3.4. Exploración de los datos**

Con el propósito de comprender la estructura general de los datos, se llevó a cabo un análisis estadístico descriptivo univariado y bivariado. Este análisis abarcó

estudios de resúmenes numéricos y gráficos relacionados con los días de morosidad y las columnas objeto de investigación.

Se emplearon funciones específicas de la librería `ggplot2`, como `geom_histogram()` y `geom_bar()`, para visualizar la distribución de los datos y comprender su dinámica de dispersión. Adicionalmente, se utilizó la función `summary()` para obtener un análisis estadístico descriptivo.

### **3.3.4. Preparación de los datos**

La finalidad de esta fase es seleccionar, limpiar y generar conjuntos de datos correctos, organizados y preparados para la fase de modelado. Se implementarán diversas técnicas para garantizar la calidad y coherencia de los datos, preparándolos de manera óptima para el análisis y construcción de modelos.

#### **3.3.4.1. Selección de datos**

En el proceso de selección de datos, se llevaron a cabo dos enfoques esenciales: por elementos dentro de las columnas `EST_CREDITO`, `SUCURSAL` y `TERRITORIO` y por atributos o características sobre las columnas con información confidencial del socio.

Dentro de la columna `EST_CREDITO`, se seleccionó los créditos cuya deuda con la Cooperativa ha sido cancelada por sus deudores (C = cancelado), descartando los créditos que aún mantienen deuda (L = liquidado), ya que de esta manera se conoce con cuantos días de mora culminó su crédito.

Para concluir con la selección por elementos, se realizó un análisis espacio-temporal con las columnas `DIAS_MORA_AC` con `FECHA_CONCESION`, agrupadas por `SUCURSAL` y `TERRITORIO`, para conocer la distribución de créditos por sucursal y territorio según la fecha de concesión, permitiendo descartar aquellas sucursales o territorios que no generaron aportes con créditos.

Para cumplir con la selección por atributos o características, se eliminó todas aquellas columnas con información confidencial de los socios y aquellas que no aportan al modelamiento del riesgo crediticio, como `COD_SOCIO`, `NOMBRE_SOCIO`, `COD_CREDITO`, `DESTINO` y `OBSERVACION`.

#### **3.3.4.2. Limpieza de datos**

Para realizar una limpieza de datos, se observó más de cerca los problemas en los datos, profundizando el análisis en las columnas que presentaban valores faltantes notables, como tipo diferimiento de crédito (`TIPO`) y calificación del riesgo crédito (`CALIFICACION`).

En el análisis de la columna TIPO, se identificó que los créditos no sujetos a refinanciamiento o reestructuración tenían valores en blanco. Por lo tanto, se asignó una categoría denominada "Original" a los valores vacíos. En cuanto a la variable CALIFICACION, se constató que los créditos cancelados no tenían calificación de riesgo, ya que esta información solo se aplicaba a los créditos vigentes según las políticas de la cooperativa. En consecuencia, se decidió no considerar esta variable en el análisis.

### **3.3.4.3. Construcción de nuevas variables**

Se construyeron nuevas variables en base a ratios financieros, reconociendo la importancia de las variables financieras en el desarrollo del modelo de riesgo crediticio. Además, se empleó codificación dummy para las variables categóricas, según se detalla en la Tabla 2.4.

Al generar los ratios financieros, se detectaron anomalías, en el saldo neto, con valores negativos. Estos resultados no concordaban con las expectativas en una solicitud de crédito; por lo tanto, se consultó con expertos, quienes indicaron que la información de los socios se actualiza anualmente y puede ser afectada por cambios en los ingresos derivados de la actividad económica, pérdidas en los cultivos o mayores gastos. Esta dinámica generaba saldos negativos, ya que los ingresos registrados para el crédito eran mayores a los actualizados. Para no perder la información de los registros de ingresos, se implementó una imputación por regresión, ajustando un modelo lineal que relacionó los ingresos (INGRESOS) con valores de gastos (GASTOS) derivados de saldos positivos.

Por último, se ajustó la dimensionalidad de variables cualitativas, instrucción (INSTRUCCION), originalmente con cinco clases (primaria, sin estudio, secundaria, formación intermedia y universitaria), se consolidó en cuatro clases: sin estudio, primaria, secundaria y superior. Asimismo, para la variable estado civil del socio (EST\_CIVIL), se generó una nueva variable dummy llamada sociedad conyugal (SOCony), que refleja la presencia o ausencia de una sociedad conyugal del socio.

**Tabla 2.4.**  
**Variables consideradas para modelamiento del riesgo crediticio**

No	Código	Variable	Descripción	Tipo	Unidades	Detalles
<b>Variables dependientes</b>						
1	DMORA	Morosidad	Días acumulados de morosidad	Cuantitativa	Días	Morosidad acumulada
<b>Variables independientes</b>						
<b>Variables sociales</b>						
1	EDSOC	Edad	Edad del socio	Cuantitativa	Años	Edad del socio
2	GENSOC	Genero	1=Masculino, 2=Femenino	Cualitativa	1:2	Asignación del genero
3	NIVEDU	Nivel de educación	1=Básico, 2=Intermedio, 3=Avanzado	Cualitativa	1:3	Nivel de estudios realizados
4	SOCONY	Sociedad conyugal	0=No, 1=Si	Cualitativa	0:1	Presencia de sociedad conyugal del socio
5	NUMCAR	Cargas	0=Sin cargas, 1=Hasta 1 carga, 2= Mas de 2 cargas	Cualitativa	0:2	Asignación de cargas familiares
<b>Variables financieras</b>						
6	ACT	Activos	Activos totales del socio	Cuantitativa	Dólares	Activos totales del socio
7	PAS	Pasivos	Pasivos totales del socio	Cuantitativa	Dólares	Deudas totales del socio
8	ING	Ingresos	Ingresos totales del socio	Cuantitativa	Dólares	Ingresos totales del socio
9	GAST	Gastos	Gastos totales del socio	Cuantitativa	Dólares	Gastos totales del socio
10	PAT	Patrimonio	Patrimonio = Activos - Pasivos	Cuantitativa	Dólares	Recursos financieros del socio
<b>Variables crediticias</b>						
11	CANT	Cantidad	Cantidad del crédito	Cuantitativa	Dólares	Cantidad o monto del crédito
12	PLAZ	Plazo del credito	1=Menos 24 meses, 2=De 24 a 48 meses, 3=Mas de 48 meses	Cualitativa	1:3	Numero de cuotas de financiamiento del credito
13	TIPINT	Tipo de interés	1=Con ahorro, 2=Sin ahorro	Cualitativa	1:2	Tipo de interés del crédito
14	INTPAG	Intereses pagados	Interés pagado o carga financiera	Cuantitativa	Dólares	Intereses generados por el crédito
<b>Variables de ubicación</b>						
15	REG	Región geográfica	1=Costa, 2=Sierra	Cualitativa	1:2	Región perteneciente la oficina del crédito
<b>Variables actividad agrícola</b>						
16	FRECING	Frecuencia de ingresos	1=Mensual, 2=Trimestral, 3=Semestral	Cualitativa	1:3	Frecuencia de ingresos agropecuarios
17	ESTCLIM	Estación climática	1=Invierno, 2=Verano	Cualitativa	1:2	Estación climática al momento del crédito

Nota: Identificación de variables dependientes y variables independientes. Variables independientes divididas de acuerdo a la fuente u origen de la información del socio y su solicitud de crédito, como variables sociales, financieras, crediticias, de ubicación y actividad agrícola.

#### **3.3.4.4. Formato de datos**

Como paso final antes del ajuste de los modelos, se aplicó un formato tipo factor a las variables cualitativas utilizando la función `factor()`, ya que algunas técnicas requieren este formato. Al mismo tiempo, se realizó una transformación del conjunto de datos de la variable de respuesta y de otros predictores con el objetivo de cumplir con los supuestos del modelo. Dependiendo de la naturaleza específica de los datos, se utilizaron las funciones `scale()` para la estandarización o `log()` para la transformación logarítmica (sumándole 1 al argumento para evitar problemas con valores iguales a cero).

#### **3.3.5. Modelamiento**

En esta fase, se aplicaron las técnicas de minería de datos utilizadas en el proceso del modelamiento de la morosidad sobre la base de datos preparada.

##### **3.3.5.1. Selección de técnicas de modelado**

Dada la presencia de excesos de ceros en la variable de interés, *Dmora* (días de mora), se optó por utilizar técnicas de regresión. Entre las opciones consideradas, las técnicas de regresión de Poisson, Poisson inflada en cero, binomial negativa y binomial negativa inflada en cero se definieron como las más apropiadas para abordar la particular distribución de la variable en cuestión.

##### **3.3.5.2. Generación de un diseño de comprobación**

Con la función `createDataPartition()` de la librería `caret` se particionó los datos en dos conjuntos, *training* (entrenamiento) y *testing* (validación) en una proporción 70% y 30%, respectivamente. A partir de esto, se validaron los modelos generados a partir de los datos de entrenamiento.

##### **3.3.5.3. Generación de los modelos**

A partir del conjunto de entrenamiento, se llevaron a cabo procesos de optimización de los parámetros inherentes a cada una de las técnicas de regresión consideradas. Con el fin de tener una óptima predicción para el cumplimiento de pago, se utilizaron métodos que permiten evitar un posible sobre ajuste del modelo.

Para implementar el modelo de regresión de Poisson, se empleó la función `glm()` de la librería `pscl`, tomando en cuenta los predictores clave. La función `zeroinfl()` de la misma biblioteca facilitó la construcción de un modelo de regresión de Poisson inflado en cero.

Asimismo, la librería `MASS` desempeñó un papel fundamental al proporcionar la función `glm.nb()`, utilizada para ajustar el modelo de regresión binomial negativa.

Finalmente, se utilizó la función `zeroinfl()` de la biblioteca `pscl` para construir el modelo de regresión binomial negativa inflada en Cero. Esta metodología permitió explorar distintas técnicas y abordajes, asegurando la robustez y la calidad predictiva de cada modelo.

### **3.3.6. Evaluación**

En esta etapa, se examinaron los resultados de los modelos desarrollados en base a criterios de rendimiento. Es decir, se evaluó el desempeño de los modelos ajustados en relación con los objetivos específicos de la investigación.

#### **3.3.6.1. Evaluación de los resultados**

Para evaluar la capacidad predictiva del modelo y determinar si existe sobreajuste, se aplicaron diversas métricas. Entre estas medidas se incluyen el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC), donde valores más bajos indican un mejor ajuste del modelo. Además, se emplearon gráficos diagnósticos, como el gráfico de residuos y el gráfico Q-Q, para evaluar la calidad del ajuste de los modelos.

# Capítulo 4

## 4. RESULTADOS

### 4.1. Base de datos

La base de datos examinada constaba de 6262 filas y 31 columnas. De estas, 15 columnas albergaban datos discretos, mientras que las 16 restantes contenían datos continuos. El número total de observaciones fue de 10962, y la memoria utilizada para almacenar esta información fue de 2.56 MB.

Este conjunto de datos proporcionó información detallada sobre la morosidad de los socios, abarcando aspectos personales, situaciones financieras, datos crediticios, ubicación geográfica financiera y detalles agrícolas. Las diferentes variables presentes en la base de datos permitieron un análisis completo de los factores que podrían afectar la morosidad.

### 4.2. Análisis de la calidad de la base de datos

Según la Tabla 4.1, se identificaron columnas con ausencia mínima o nula de datos, representadas por las columnas "COD\_SOCIO", "ACTIVOS", "PASIVOS", "INGRESOS", "GASTOS", "SUCURSAL", "TERRITORIO", "COD\_CREDITO", "EST\_CREDITO", "FECHA\_CONCESION", "FECHA\_VENCIMIENTO", "PLAZO", "MONTO", "TASA\_INTERES", "DESTINO", "SALDO\_TOTAL", "SALDO\_VENCIDO", "DIAS\_MORA\_AC", "PIGNORACION" y "TIPO". Estas columnas exhibieron integridad completa en sus datos.

En contraste, se observó la presencia de valores faltantes en ciertas columnas, aunque en porcentajes menores al 10%. Columnas como "EDAD", "GENERO", "INSTRUCCION", "EST\_CIVIL", "N\_CARGAS", "NOMBRE\_SOCIO", "OBSERVACION", "ACTIVIDAD\_AGRO" y "GRUPO\_AGRO" presentaron ausencia de datos, pero los porcentajes no superaron el umbral del 10%.

Por otro lado, la variable "CALIFICACION" resaltó como la más afectada, con valores faltantes que superaban el 50%. Este alto porcentaje podría tener un impacto significativo en análisis estadísticos y modelados que dependan de esta columna. Por tanto, se optó por omitirla en el análisis para evitar posibles distorsiones en los resultados o conclusiones derivadas de su inclusión (ver Fig. 4.1).

**Tabla 4.1. Proporción de valores perdidos por variable**

NC	Columna	Tipo	N, perdidos	%, Columna	N, valores
1	COD_SOCIO	Numérico	0	0.00%	5338
2	EDAD	Numérico	30	0.48%	77
3	GENERO	Carácter	30	0.48%	3
4	INSTRUCCION	Carácter	30	0.48%	7
5	EST_CIVIL	Carácter	30	0.48%	6
6	N_CARGAS	Numérico	30	0.48%	10
7	ACTIVOS	Numérico	0	0.00%	2847
8	PASIVOS	Numérico	0	0.00%	3292
9	INGRESOS	Numérico	0	0.00%	1377
10	GASTOS	Numérico	0	0.00%	2842
11	NOMBRE_SOCIO	Carácter	29	0.46%	5312
12	SUCURSAL	Carácter	0	0.00%	30
13	TERRITORIO	Carácter	0	0.00%	9
14	COD_CREDITO	Numérico	0	0.00%	6261
15	EST_CREDITO	Carácter	0	0.00%	2
16	FECHA_CONCESION	Carácter	0	0.00%	2429
17	FECHA_EVNCIMIENTO	Carácter	0	0.00%	3012
18	PLAZO	Numérico	0	0.00%	61
19	MONTO	Numérico	0	0.00%	285
20	TASA_INTERES	Numérico	0	0.00%	11
21	DESTINO	Carácter	0	0.00%	14
22	SALDO_TOTAL	Numérico	0	0.00%	1649
23	SALDO_VENCIDO	Numérico	0	0.00%	89
24	CALIFICACION	Carácter	4136	66.05%	10
25	DIAS_MORA_AC	Numérico	0	0.00%	968
26	PIGNORACION	Numérico	0	0.00%	70
27	INTERES_PAGADO	Numérico	0	0.00%	6001
28	OBSERVACION	Carácter	5	0.08%	4912
29	TIPO	Carácter	0	0.00%	5
30	ACTIVIDAD_AGRO	Carácter	587	9.37%	37
31	GRUPO_AGRO	Carácter	587	9.37%	4

Nota: NC=Ubicación de columna en la base de datos. N, valores = número de valores únicos por columna.

En la Fig. 4.1 se observó que las columnas “EDAD”, “GÉNERO”, “INSTRUCCIÓN”, “EST\_CIVIL”, “N\_CARGAS” y “NOMBRE\_SOCIO” presentaban coincidencias en los lugares de valores faltantes, cada una registrando solamente 30 valores perdidos (0.48%) por columna. Al considerar el tipo de dato de las columnas, se sugiere la posibilidad de que estos valores faltantes sean no aleatorios. Este hallazgo motivó la decisión de eliminar dichos valores faltantes.

Se identificó que los valores faltantes en las columnas “ACTIVIDAD\_AGRO” y “GRUPO\_AGRO” (ver Fig. 4.1) presentan similitudes tanto en su ubicación como en su proporción. Dado que estas variables son de tipo carácter y el origen de los datos no es claro respecto a si estos valores faltantes son aleatorios o no. Por lo tanto, se procedió a eliminar aquellos valores faltantes.



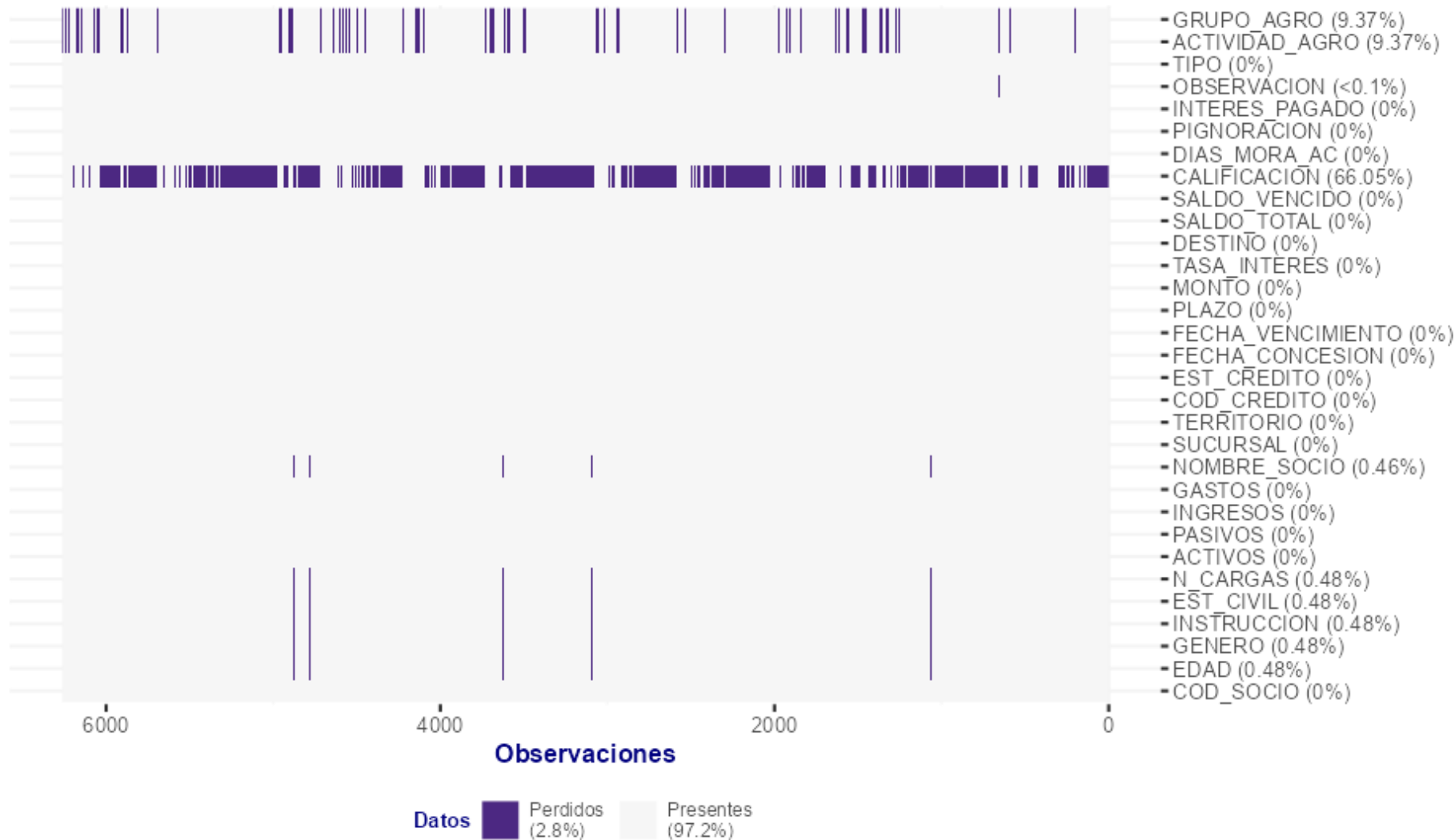


Fig. 4.1. Proporción y ubicación de valores perdidos (NA) en la base datos.

El análisis de valores duplicados en la Tabla 4.2, demostró resultados significativos. Se observó una similitud notable en la cantidad de duplicados entre las columnas "COD\_SOCIO" y "NOMBRE\_SOCIO", con 910 registros duplicados. Estos resultados sugieren la posibilidad de que ciertos socios hayan adquirido múltiples créditos productivos a lo largo del tiempo, indicando una repetida participación en el sistema crediticio.

El descubrimiento de duplicados en la columna 'COD\_CREDITO', es inesperado, ya que la presencia de múltiples créditos con condiciones idénticas no debería ocurrir. Ante esta situación, se ha tomado la decisión de eliminar estas filas duplicadas con el fin de prevenir posibles complicaciones derivadas de casos de duplicación durante el análisis de los datos.

**Tabla 4.2.**  
**Proporciones y numero de valores duplicados por columna.**

NC	Columna	N, Duplicados	%, Columna	%, Total
1	COD_SOCIO	910	16.11%	0.54%
2	EDAD	5573	98.65%	3.29%
3	GENERO	5647	99.96%	3.33%
4	INSTRUCCION	5643	99.89%	3.33%
5	EST_CIVIL	5644	99.91%	3.33%
6	N_CARGAS	5640	99.84%	3.33%
7	ACTIVOS	3112	55.09%	1.84%
8	PASIVOS	2694	47.69%	1.59%
9	INGRESOS	4439	78.58%	2.62%
10	GASTOS	3101	54.89%	1.83%
11	NOMBRE_SOCIO	910	16.11%	0.54%
12	SUCURSAL	5621	99.50%	3.32%
13	TERRITORIO	5640	99.84%	3.33%
14	COD_CREDITO	1	0.02%	<0.01%
15	EST_CREDITO	5647	99.96%	3.33%
16	FECHA_CONCESION	3294	58.31%	1.94%
17	FECHA_VENCIMIENTO	2800	49.57%	1.65%
18	PLAZO	5591	98.97%	3.30%
19	MONTO	5375	95.15%	3.17%
20	TASA_INTERES	5640	99.84%	3.33%
21	DESTINO	5637	99.79%	3.33%
22	SALDO_TOTAL	4410	78.07%	2.60%
23	SALDO_VENCIDO	5590	98.96%	3.30%
24	DIAS_MORA_AC	4774	84.51%	2.82%
25	PIGNORACION	5583	98.83%	3.29%
26	INTERES_PAGADO	199	3.52%	0.12%
27	OBSERVACION	1304	23.08%	0.77%
28	TIPO	5644	99.91%	3.33%
29	ACTIVIDAD_AGRO	5613	99.36%	3.31%
30	GRUPO_AGRO	5646	99.95%	3.33%

Nota: NC = Posición del número de columna en el conjunto de datos. N, Duplicados = Numero de valores duplicados por columna.

En la Tabla 4.3 se observó que la mayoría de las columnas tiene valores únicos que se alinean con lo esperado y las categorías predefinidas. Sin embargo, se detectó la presencia de valores como '0' en la columna 'ACTIVIDAD\_AGRO'. Esto sugiere la posibilidad de ausencia de información o la representación intencional de ciertos datos. Se consideró necesario realizar una verificación adicional para determinar si estos valores indican datos faltantes o tenían una interpretación específica dentro del contexto del conjunto de datos.

**Tabla 4.3.**  
**Valores únicos para columnas categóricas.**

Columna	N, Valores	Valores únicos
GENERO	2	FEMENINO, MASCULINO
INSTRUCCION	6	SIN ESTUDIOS, PRIMARIA, SECUNDARIA, UNIVERSITARIA, FORMACION INTERMEDIA (TECNICA - TECNOLOGIA), NO DEFINIDO
EST_CIVIL	5	CASADO/A, SOLTERO/A, DIVORCIADO/A, VIUDO/A, UNION LIBRE
*NOMBRE_SOCIO	4739	-
SUCURSAL	28	AZOGUES, MILAGRO, CAÑAR, CHORDELEG, CUENCA, DURAN, GUALACEO, GUALAQUIZA, GUAYAQUIL, HUAMBOYA, MACAS, MACHALA, MACHALA CENTRO, NABÓN, NARANJAL, OÑA, PALMAS, PASAJE, PATAMARCA, PAUTE, PONCE ENRIQUEZ, SANTA ROSA, SARAGURO, SEVILLA DE ORO, SIGSIG, SUCUA, TRONCAL, ZARUMA
TERRITORIO	9	CAÑAR, COSTA DOS, PAUTE, CUENCA, ORIENTE, COSTA UNO, COSTA TRES, JUBONES DOS, JUBONES UNO
EST_CREDITO	2	C, L
DESTINO	12	CT, AFTT, NP, AFTVL, MCM, AFTO, AS, AFTE, CAFT, AFIA, CAFI, AFTVP
*OBSERVACION	4345	-
TIPO	5	ORIGINAL, REPRESTAMO, REFINANCIADO, NOVADO, REESTRUCTURADO
ACTIVIDAD_AGRO	36	GANADERIA, PECUARIOS, NO AGROPECUARIAS, TOMATE, CICLO CORTO, AVICULTURA, PORCICULTURA, COBAYO, PAPA, 0, HORTICOLAS, MORA, ARROZ, CACAO, BANANO, FRESA, ORQUIDEAS, TOMATE ARBOL, SOYA, MAIZ, BABACO, MADERA, PITAHAYA, MARACUYA, CEBOLLA, FRUTALES, HIGUERA, AGUACATE, PIMIENTO, LIMON, MENSUAL, PAPAYA, CAFÉ, QUINUA, PLATANO, CAÑA DE AZUCAR
GRUPO_AGRO	3	PECUARIOS, NO AGROPECUARIOS, AGRICOLAS

Nota: Descripción del número de valores únicos por columna. CT = Capital de trabajo, AFTT = Activos fijos tangibles: terrenos, edificios y construcción de infraestructura, para fines productivos y comerciales., NP = No productivas (para operaciones inmobiliarias, vivienda de interés público y crédito educativo, no otorgado a microempresarios)., AFTVL = Activos fijos tangibles: vehículos livianos para fines productivos y comerciales., MCM = Microcrédito otorgado para consumo de microempresarios., AFTO = Activos fijos tangibles: otros activos fijos tangibles que no constituyen bienes de capital, para fines productivos y comerciales, AS = Adquisición de servicios., AFTE = Activos fijos tangibles: equipos, maquinaria y otros bienes de capital, a excepción de vehículos, para fines productivos y comerciales., CAFT = Compra de activos fijos tangibles., AFIA = Activos fijos intangibles: adquisición de franquicias, marcas, pago de regalías, licencias y otros activos fijos intangibles., CAFI = Compra de activos fijos intangibles, AFTVP = Activos fijos tangibles: vehículos pesados para fines productivos y comerciales.

\*No se muestran los valores debido a la extensión de los datos o a sus características específicas.

En la Tabla 4.4, las columnas COD\_CREDITO y COD\_SOCIO presentaron 5648 y 4734 valores únicos, respectivamente, lo cual indica una diversidad extensa en los identificadores de socios.

La variable "PLAZO" contenía valores entre 0 y 504, siendo el cero común en créditos a corto plazo; sin embargo, el máximo de 504 demandaba verificación para asegurar su coherencia. Al observar "N\_CARGAS", se identificó un rango de 0 a 9, y el valor de cero planteaba la posibilidad de personas que no dependían económicamente del solicitante.

Por su parte, "DIAS\_MORA\_AC", con un máximo de 20079 días, indicaba posibles valores incoherentes en la entrada de datos o casos donde el socio había entrado a procesos judiciales. El valor mínimo de cero en "INGRESOS" y "GASTOS" generaba interrogantes sobre la fiabilidad de esos datos, resultando inusual en el contexto de créditos. Se requería una evaluación más detallada para confirmar si estos ceros eran representativos de los datos reales o si podrían ser errores o valores faltantes codificados como cero.

Las variables "ACTIVOS" y "PASIVOS" presentaban valores mínimos de cero, que podrían ser legítimos en ciertos contextos financieros; no obstante, se necesitaba una evaluación detallada para confirmar su representatividad.

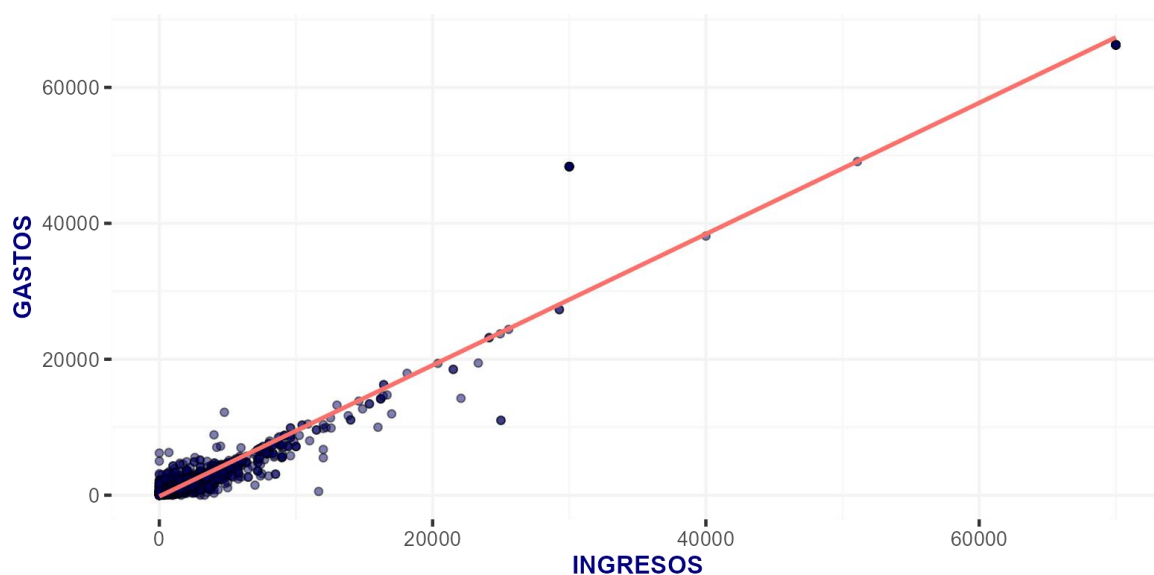
Asimismo, las extremidades de SALDO\_TOTAL, SALDO\_VENCIDO, PIGNORACION e INTERES\_PAGADO necesitaban verificación para garantizar su coherencia y fiabilidad dentro del conjunto de datos.

**Tabla 4.4.**  
**Valores únicos para columnas numéricas.**

<b>Columna</b>	<b>N, valores únicos</b>	<b>Mínimo</b>	<b>Máximo</b>
<i>COD_SOCIO</i>	4734	10368	793077
EDAD	76	18	99
N_CARGAS	9	0	9
ACTIVOS	2537	0	2710000
PASIVOS	2955	0	334872.91
INGRESOS	1210	0	70000
GASTOS	2548	0	66260
<i>COD_CREDITO</i>	5648	1386	915643
PLAZO	58	0	504
MONTO	274	170	60000
TASA_INTERES	9	11.22	16
SALDO_TOTAL	1239	0	43297.76
SALDO_VENCIDO	59	0	6516.28
<i>DIAS_MORA_AC</i>	878	0	20079
PIGNORACION	66	0	1350
<i>INTERES_PAGADO</i>	5450	0	21025.55

Nota: Descripción del número de valores únicos, mínimo y máximo por columna.

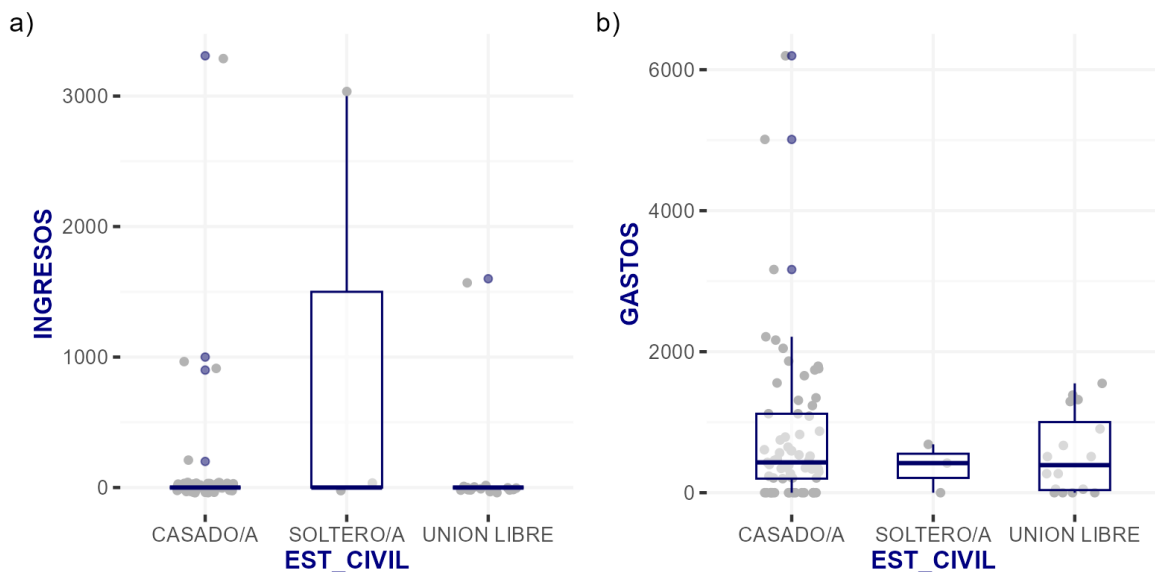
Para llevar a cabo la imputación de datos, se examinó la relación entre ingresos y gastos, una consideración crucial en las solicitudes de crédito para determinar la capacidad de pago [60]. Puesto que, un ingreso superior a los gastos puede dar lugar a una relación gastos-ingresos más baja, lo cual se considera favorable. Es decir, cuando los ingresos son mayores que los gastos, la relación es más ventajosa desde el punto de vista financiero [61]. Estos conceptos se visualizan en la Fig. 4.2, donde se representa la relación entre ingresos y gastos.



**Fig. 4.2. Diagrama de dispersión entre INGRESOS y GASTOS**

Nota: La línea roja corresponde al modelo de ajuste lineal que representa la relación estimada entre las variables de Ingresos y Gastos, según el análisis de regresión aplicado a los datos.

El análisis de los valores con cero en las columnas "GASTOS" e "INGRESOS", según el estado civil de los socios ("CASADO/A", "SOLTERO/A", "DIVORCIADO/A", "VIUDO/A", "UNION LIBRE"), reveló una asociación marcada en las solicitudes de socios en relaciones conyugales ("CASADO/A" o "UNION LIBRE"), con menor presencia en socios "SOLTERO/A". Estos valores nulos podrían vincularse con el registro de ingresos o gastos del solicitante del crédito. Es decir; si el solicitante no reporta ingresos, estos podrían ser justificados por su cónyuge, lo que puede causar error al registrar e interpretar los datos (ver Fig. 4.3).



**Fig. 4.3. Boxplot de INGRESOS y GASTOS según EST\_CIVIL**

Nota: a) Boxplot de INGRESOS. b) Boxplot de GASTOS

Al analizar la Tabla 4.5, se observa que la “EDAD” de los registros oscila entre 18 y 99 años, mostrando una distribución que se concentra en rangos menores a los 56 años para el 75% de los casos. Las columnas financieras, como "ACTIVOS" y "PASIVOS", presentan una marcada variabilidad, con rangos significativos entre valores mínimos y máximos. Las columnas “INGRESOS” y “GASTOS” tienen una dispersión considerable, con la mayoría de los gastos inferiores a 1095 y una proporción de ingresos por debajo de los 500. Las columnas relacionadas con saldos (“SALDO\_TOTAL” y “SALDO\_VENCIDO”) reflejan una presencia notable de valores cero, sugiriendo que una gran parte de los registros no tienen saldos pendientes, vencidos o pignoraciones. Además, "DIAS\_MORA\_AC" muestra una amplia gama de valores, desde cero hasta 20079 días, mientras que "INTERES\_PAGADO" varía significativamente entre cero y 21026, mostrando una dispersión considerable en el conjunto de datos.

Además, se observa que no hay valores que parezcan incoherentes o extraordinarios a primera vista. Sin embargo, la presencia de muchos valores cero en "SALDO\_TOTAL", "SALDO\_VENCIDO" y "PIGNORACION" es notable y podría indicar una tendencia específica en estos campos. La amplia gama de días en mora en "DIAS\_MORA\_AC" también llama la atención, particularmente aquellos registros con un número extremadamente alto de días. Estos podrían ser considerados atípicos y necesitarían una revisión más detallada para determinar si son errores de ingreso o si representan situaciones genuinas, pero poco comunes.

**Tabla 4.5.**  
**Resumen estadístico de variables cuantitativas.**

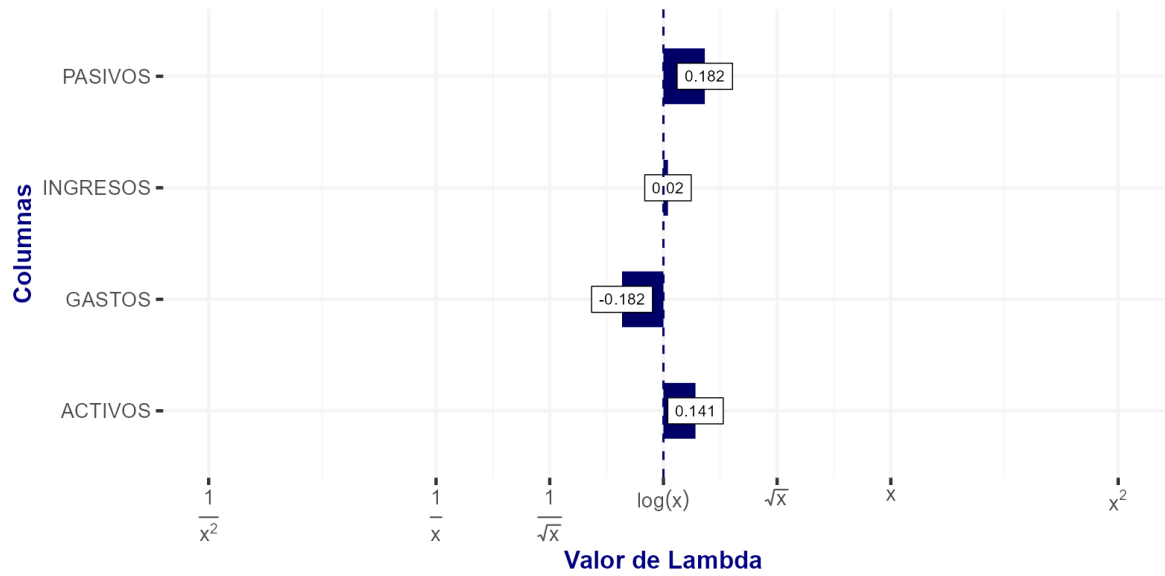
<b>Columnas</b>	<b>N</b>	<b>Mínimo</b>	<b>Q1</b>	<b>Mediana</b>	<b>Q3</b>	<b>Máximo</b>
ACTIVOS	5358	5	14301	40000	84007	2710000
DIAS_MORA_AC	5358	0	0	12	84	20079
EDAD	5358	18	33	44	57	99
GASTOS	5358	5	370	610	1095	66260
INGRESOS	5358	1	500	800	1350	70000
INTERES_PAGADO	5358	0	192	501	1173	21026
MONTO	5358	170	2300	5000	8000	50000
N_CARGAS	5358	0	0	1	2	9
PASIVOS	5358	0	444	4513	13017	334873
PIGNORACION	5358	0	0	0	0	1350
PLAZO	5358	1	24	36	60	84
SALDO_TOTAL	5358	0	0	0	355	43298
SALDO_VENCIDO	5358	0	0	0	0	6516
TASA_INTERES	5358	11	13	13	15	16

Nota: Resumen estadístico por columna. N = número de datos. Q1 = primer cuartil (percentil 25%). Q3 = tercer cuartil (percentil 75%).

Con el propósito de mejorar la simetría y ajustar desigualdades en las distribuciones de datos, se calcularon los valores óptimos de lambda mediante las transformaciones Box-Cox en diversas columnas del conjunto de datos, tal como se ilustra en la Fig. 4.4.

En la columna "PASIVOS", se identificó un valor de lambda de 0.182, indicando la aplicación de una transformación logarítmica para corregir la asimetría positiva y lograr una distribución más simétrica. En cuanto a "INGRESOS", el valor de lambda cercano a cero (0.02) sugiere la posibilidad de una transformación logarítmica, aunque se aconseja realizar una evaluación adicional para determinar si son necesarios ajustes adicionales, considerando la posible normalidad inicial de la distribución.

Para la columna "GASTOS", el lambda de -0.182, al ser un valor negativo y mayor que -0.5, aun sugiere una inclinación hacia una transformación logarítmica. Respecto a "ACTIVOS", con un lambda de 0.141, siendo positivo y menor que 0.5, sugiere la aplicación de una transformación, como la raíz cuadrada o logarítmica, para normalizar la distribución de datos.



**Fig. 4.4. Valores óptimos de lambda obtenidos por transformaciones Box-Cox.**

Nota: Los valores de lambda representan los óptimos obtenidos para cada columna en el análisis de transformación Box-Cox. Estos valores óptimos se utilizan en técnicas de transformación para mejorar la distribución y homogeneizar la varianza de los datos en cada categoría, lo que puede ser crucial en análisis estadísticos y modelado predictivo.

La



**Tabla 4.6** detalla cómo distintas transformaciones afectan la presencia de valores atípicos y normales en las columnas evaluadas. En "ACTIVOS", la transformación logarítmica destaca con solo un 0.82% de valores atípicos, sugiriendo mayor consistencia. Para "GASTOS" y "INGRESOS", el logaritmo también muestra bajos porcentajes de valores atípicos (2.56% y 3.71%, respectivamente), indicando distribuciones más estables. Sin embargo, en "PASIVOS", el logaritmo presenta el mayor porcentaje de valores atípicos (21.30%), indicando la necesidad de una evaluación cuidadosa al interpretar los resultados asociados con esta transformación.

Por lo tanto, la transformación logarítmica parece ser la opción más robusta para normalizar los datos en "ACTIVOS", "GASTOS" y "INGRESOS". Para "PASIVOS", la elección depende de la tolerancia a valores extremos y la interpretación contextual.

**Tabla 4.6.**  
**Valores atípicos y normales según transformaciones Box-Cox.**

Columnas	Transformación Box-Cox					
	$\sqrt{(x+1)}$		$\log(x+1)$		$1/\sqrt{(x+1)}$	
	Atípico	Normal	Atípico	Normal	Atípico	Normal
ACTIVOS	2.78%	97.22%	<b>0.82%</b>	99.18%	7.50%	92.50%
GASTOS	6.25%	93.75%	2.56%	97.44%	<b>1.77%</b>	98.23%
INGRESOS	6.94%	93.06%	3.71%	96.29%	<b>1.87%</b>	98.13%
PASIVOS	<b>1.42%</b>	98.58%	21.30%	78.70%	22.40%	77.60%

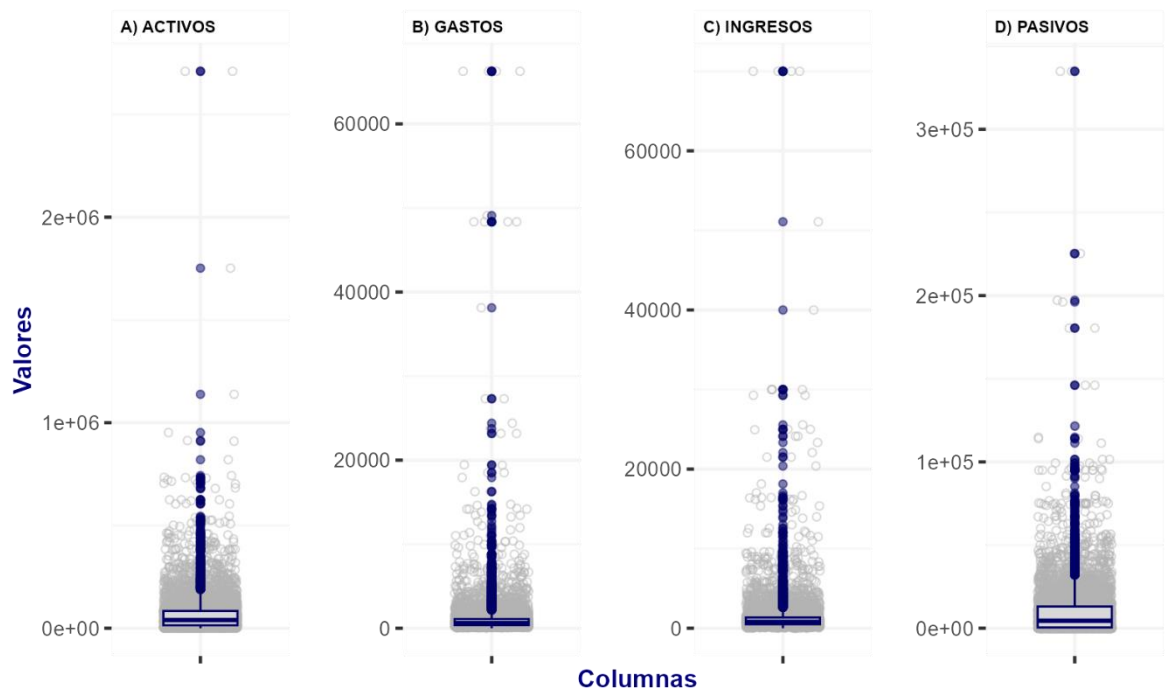
Nota: La tabla muestra los porcentajes de valores atípicos y valores no atípicos para las variables financieras ('ACTIVOS', 'GASTOS', 'INGRESOS', 'PASIVOS') bajo distintas transformaciones Box-Cox. Los valores atípicos fueron identificados utilizando el criterio de encontrarse a una distancia del primer cuartil y del tercer cuartil superior a 1.5 veces el rango intercuartílico.

En la Fig. 4.5, se observaron valores que indicaban que la mayoría de los datos se agrupaban alrededor del valor cero, sugiriendo que la mayoría de los socios tenían valores bajos en activos, gastos, ingresos y pasivos. A pesar de esta concentración en cero, la presencia de outliers señaló que algunos socios presentaban valores significativamente más altos en estas variables.

Se pudo notar que la caja de los boxplot de los valores originales, era estrecha, lo que sugiere que la dispersión de la mayoría de los datos estaba concentrada en un rango relativamente pequeño alrededor de la mediana. En un boxplot, la caja representa el rango intercuartil (IQR), que abarca desde el primer cuartil (Q1) hasta el tercer cuartil (Q3). Cuanto más estrecha sea la caja, mayor será la concentración de los datos dentro de este rango intercuartil.

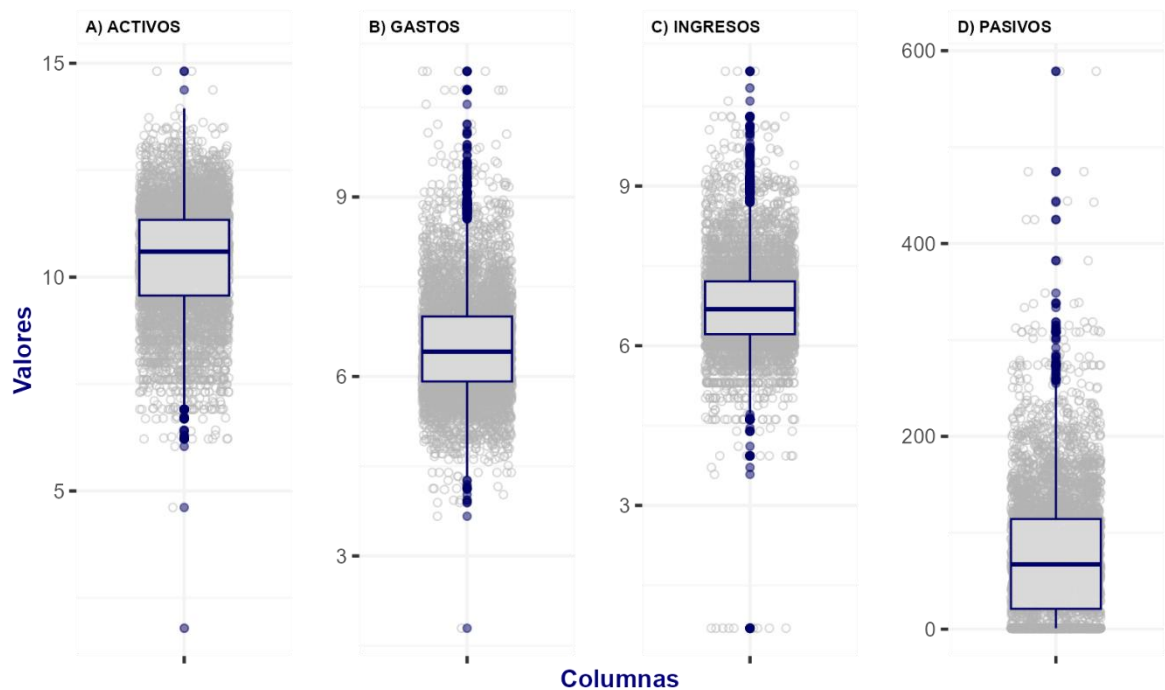
Para mejorar la distribución de los datos, se decidió llevar a cabo una transformación según el análisis de la Tabla 4.6. Esta transformación mostró una distribución mejorada, con los valores concentrados en la parte baja del Boxplot, aunque los outliers se vieron afectados por esta transformación.

La transformación de los datos logró estabilizar la varianza, considerando que los datos originales presentaban una gran variabilidad. Aunque no se logró reducir la presencia de valores outliers, la mejora en la distribución de los datos facilita análisis posteriores y contribuye a una mejor interpretación de los resultados.



**Fig. 4.5. Boxplot de columnas financieras.**

Nota: Análisis basado en valores reales de las variables financieras. A) Activos. B) Gastos. C) Ingresos. D) Pasivos



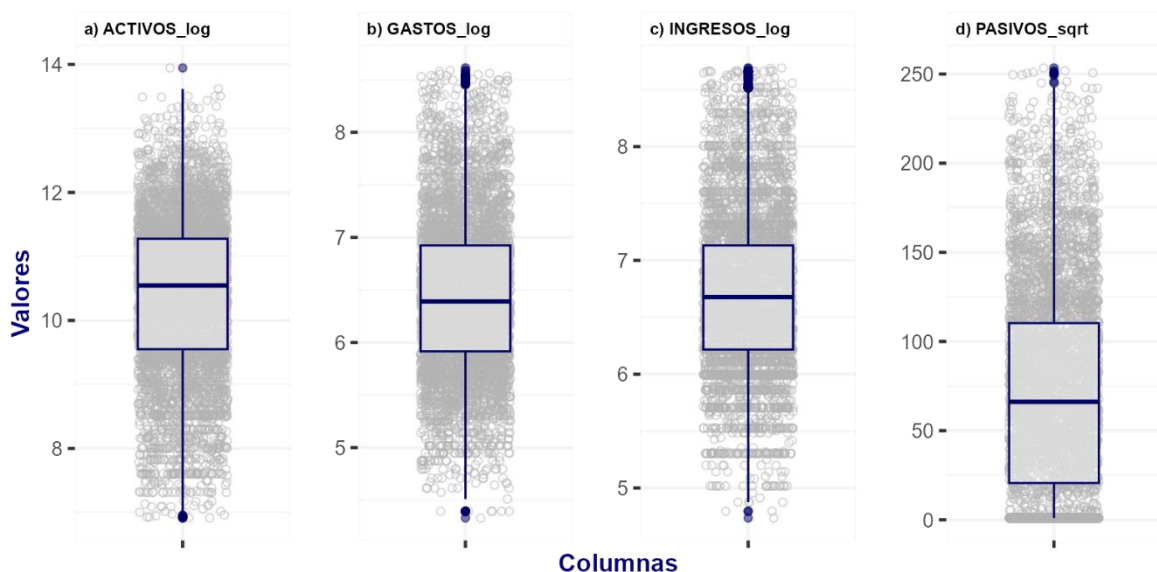
**Fig. 4.6. Boxplot de columnas financieras con transformación Box-Cox.**

Nota: Transformación Box-Cox aplicada variables financieras. A) Activos. B) Gastos. C) Ingresos. D) Pasivos

Durante el análisis de outliers, se llevaron a cabo ajustes de acuerdo con las políticas establecidas por la institución financiera, especialmente en lo que respecta a evitar la presencia de valores cero en las columnas de ingresos y gastos. Para abordar los outliers en las columnas de pasivos y activos, se aplicó el método del rango intercuartílico con el objetivo de preservar la integridad en la interpretación de los resultados del análisis.

En la Fig. 4.7, se evidenció que, tras la transformación y eliminación de outliers, los boxplots revelaron una mayor concentración de los datos. Las cajas intercuartiles se volvieron más estrechas, indicando una reducción en la variabilidad dentro del rango intercuartílico. La disminución en la presencia de valores atípicos contribuyó a la reducción de la dispersión en los datos.

Durante el análisis, se observó que al eliminar outliers de una columna específica, otras columnas presentaron nuevos outliers o experimentaron cambios en sus distribuciones. Se consideró la posibilidad de que esto ocurriera debido a relaciones o correlaciones entre las variables. La presencia o eliminación de outliers en una columna pudo influir en la distribución o estadísticas de otras variables en el conjunto de datos [62].

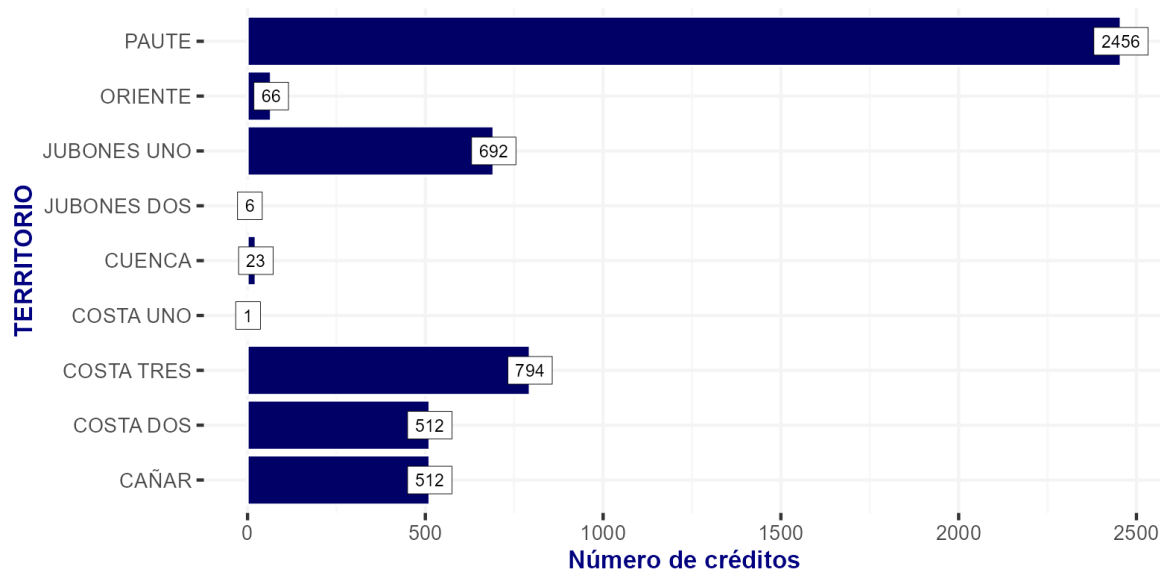


**Fig. 4.7. Boxplot de columnas financieras después eliminación de outliers.**

Nota: Transformación Box-Cox aplicada variables financieras. A) Activos. B) Gastos. C) Ingresos. D) Pasivos

Los datos sobre el número de créditos otorgados por territorios revelaron que algunos, como PAUTE, JUBONES UNO, COSTA TRES y CAÑAR, tenían muchos créditos, entre 512 y 2456. Pero otros, como COSTA UNO con 1 crédito, JUBONES DOS con 6 créditos, ORIENTE con 66 créditos y CUENCA con 23 créditos, tenían cifras más bajas.

Para mejorar el modelo de regresión y hacerlo más preciso, se decidió quitar los territorios con menos de 100 créditos. Esta elección se basa en pensar que los territorios con pocos créditos podrían no influir mucho en el modelo. Al quitar estos territorios al dividir los datos en conjuntos de entrenamiento y prueba, se espera que el modelo se ajuste mejor y pueda prever con mayor precisión el comportamiento en territorios más representativos y significativos en términos crediticios.

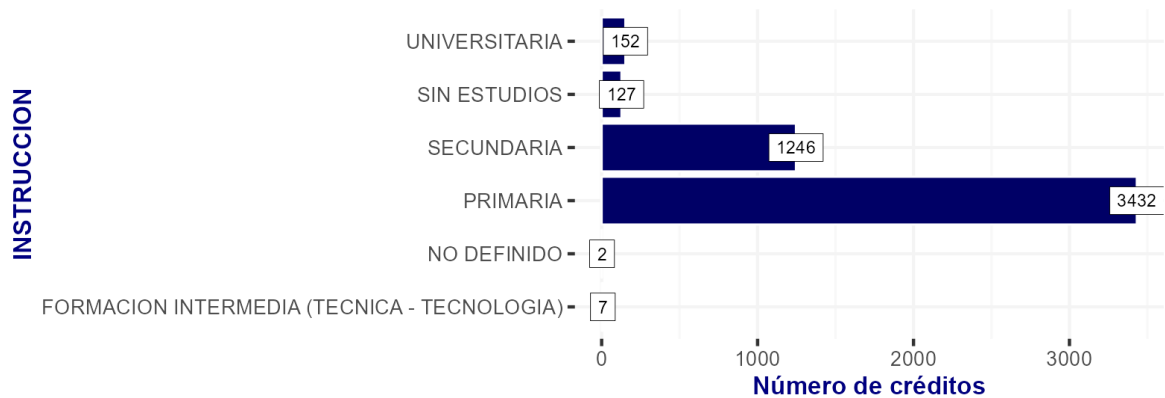


**Fig. 4.8. Numero de creditos por territorio.**

Nota: Distribución del número de creditos por territorios.

En el análisis del nivel de instrucción, se observaron variaciones en la distribución, donde algunos niveles obtuvieron cifras considerables como PRIMARIA, “SECUNDARIA” y “UNIVERSITARIA”, con valores entre 1246 y 3432. Sin embargo, se identificaron niveles con una cantidad muy reducida de personas, específicamente “FORMACIÓN INTERMEDIA” con 7 personas, “NO DEFINIDO” con 2 personas y “SIN ESTUDIOS” con 127 personas.

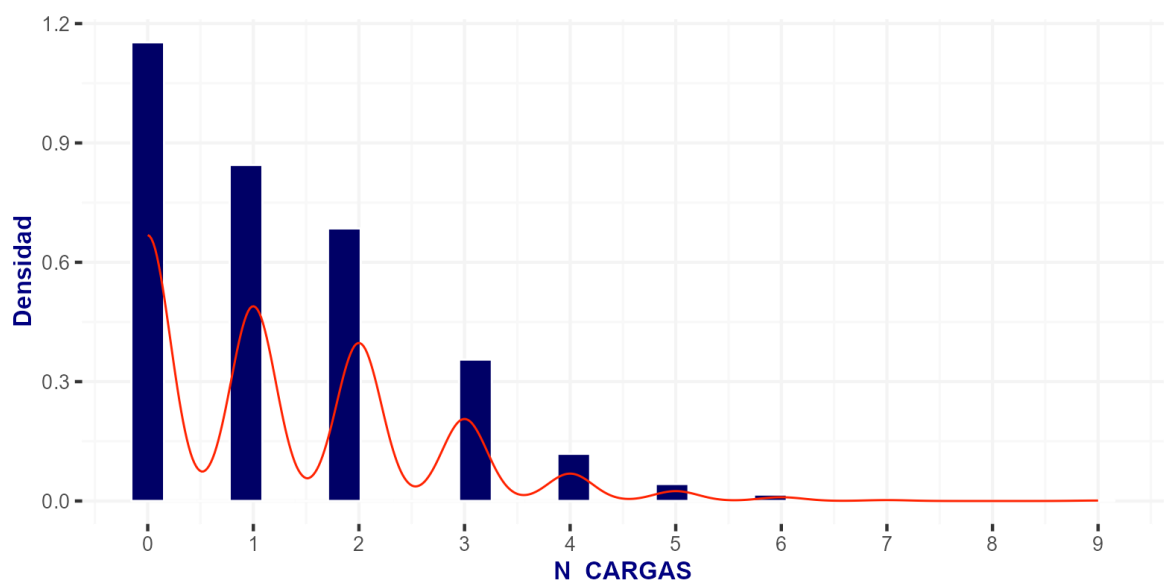
Con el objetivo de mejorar la robustez del modelo de regresión y evitar problemas potenciales durante el proceso de modelado, se optó por eliminar los niveles de instrucción, especialmente aquellos con cifras muy bajas, como NO DEFINIDO. Para mantener la información relevante de las otras columnas y su importancia en la modelación del riesgo crediticio, se procedió a reagrupar los niveles restantes en una columna denominada NIV\_INSTRUCCION, que constó de tres niveles: “BASICO”, “INTERMEDIO” y “AVANZADO”, permitiendo así conservar la información esencial para el análisis del riesgo crediticio.



**Fig. 4.9. Distribución de los niveles de INSTRUCCION.**

Nota: Distribución del número de niveles de INSTRUCCION.

La distribución del número de cargas, representada por el histograma de la **¡Error! No se encuentra el origen de la referencia.** y su línea de densidad, exhibió oscilaciones con picos que coincidían con los valores medios de las barras del histograma. Se identificó que las cargas representaban niveles específicos, abarcando desde cero hasta nueve. Debido a esta naturaleza discreta, se tomó la decisión de tratar la variable como categórica y se agrupó en tres categorías: "SIN CARGAS", "HASTA 1 CARGA" y "MÁS DE 2 CARGAS". Esta elección permitió realizar un análisis de frecuencia más detallado, facilitando la interpretación de los resultados al examinar la distribución específica de cada categoría.

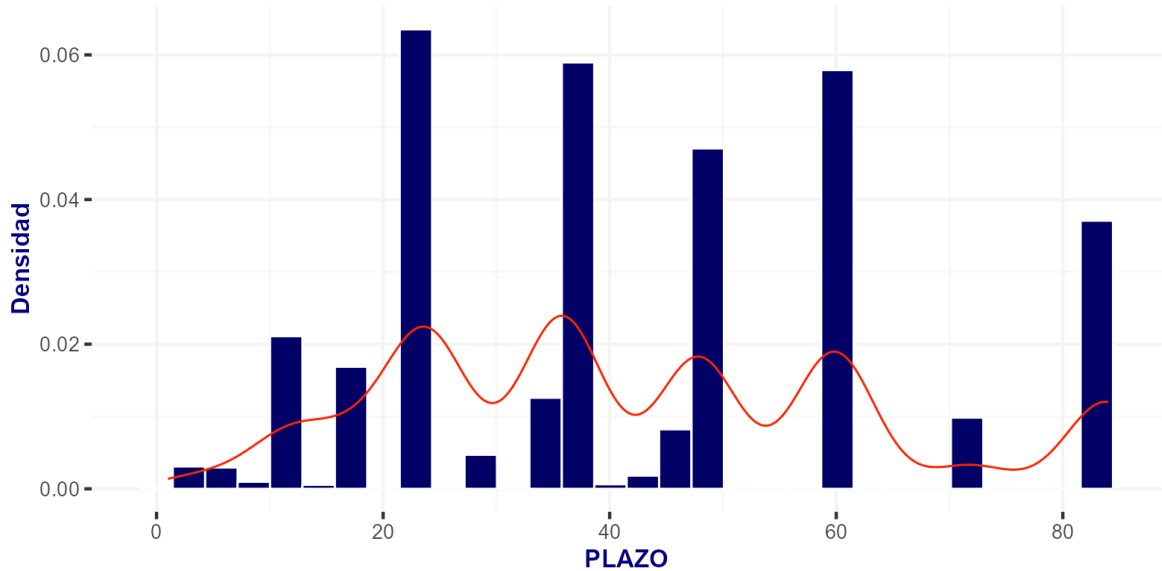


**Fig. 4.10. Histograma y curva de densidad para N\_CARGAS.**

Nota: Distribución del número de cargas (N\_CARGAS) representada por histograma y curva de densidad.

Se observó que el histograma de la columna "PLAZO" en la Fig. 4.11 no presentó una distribución que siguiera claramente un patrón específico. En lugar de una forma de distribución definida, se identificaron niveles o agrupaciones distintas. Por lo tanto, se tomó la decisión de categorizar la variable en tres grupos, creando

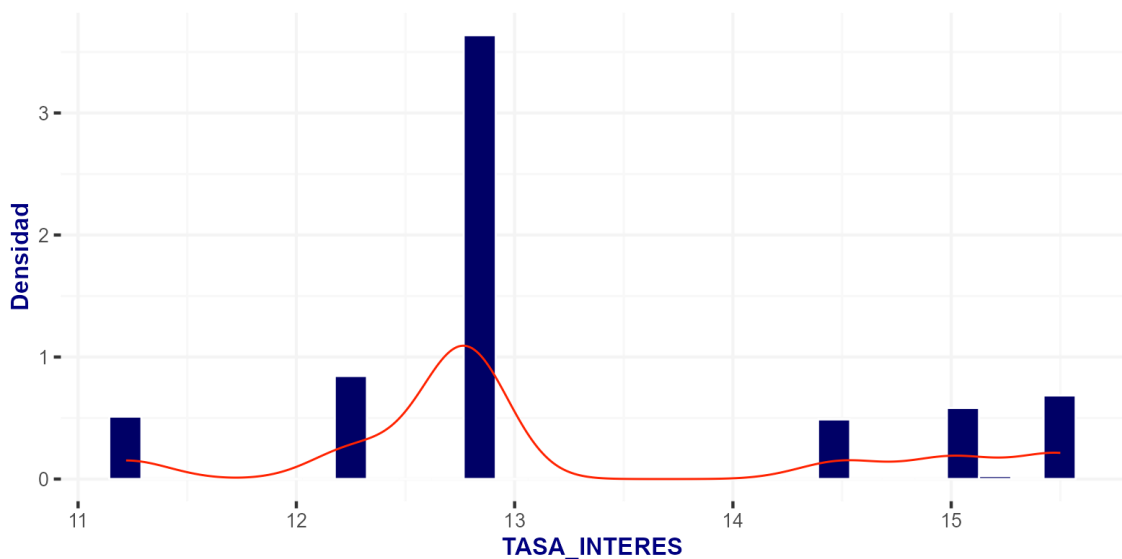
una nueva columna llamada NIV\_PLAZO, con las categorías "MENOS DE 24 MESES", "DE 24 a 48 MESES" y "MAS DE 48 MESES". Este enfoque de categorización simplifica la interpretación y el análisis de la variable en cuestión.



**Fig. 4.11. Histograma y curva de densidad para PLAZO.**

Nota: Distribución de PLAZO, representada por histograma y curva de densidad.

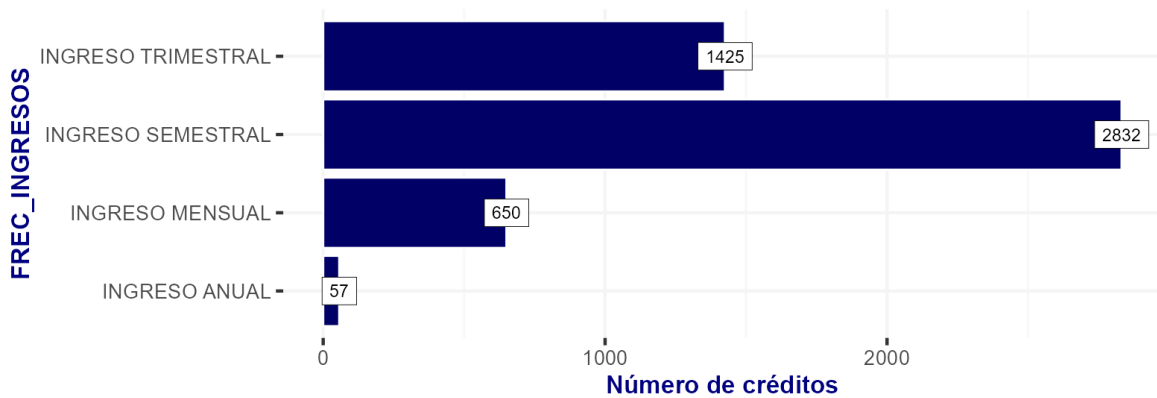
La Fig. 4.12, que presentaba el histograma y la línea de densidad, reveló que la tasa de interés mostró una concentración significativa cerca de 13. La mayor parte de los datos se agruparon alrededor de este valor, indicando una distribución centrada en torno a él. Además, se observó que la dispersión de los datos no era amplia. Siguiendo la lógica aplicada en el análisis de las columnas previas, se decidió categorizar la variable en una nueva columna llamada "NIV\_INTERES", con dos niveles: "SIN AHORRO" y "CON AHORRO".



**Fig. 4.12. Histograma y curva de densidad para TASA\_INTERES.**

Nota: Distribución de TASA\_INTERES, representada por histograma y curva de densidad.

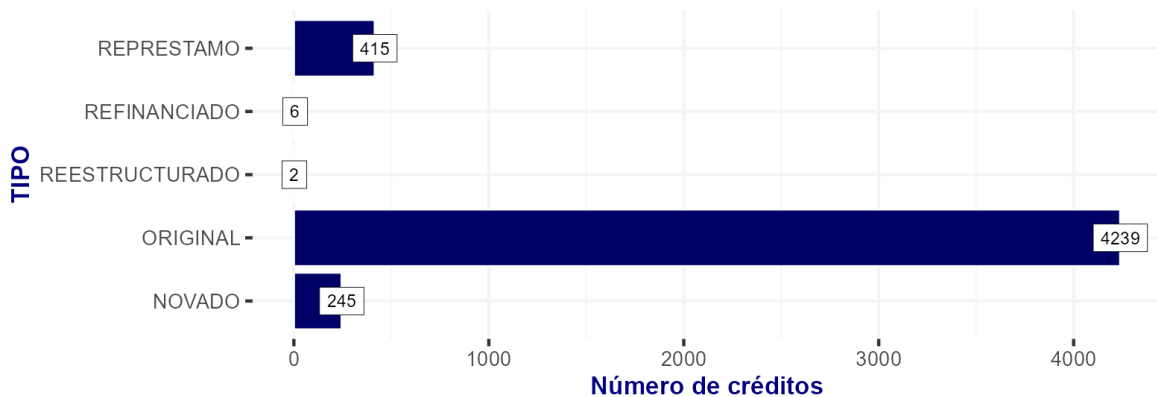
En el análisis de la columna "FREC\_INGRESOS", se destacaron las categorías más frecuentes: "INGRESO TRIMESTRAL" con un valor de 1425, "INGRESO SEMESTRAL" con 2832 y "INGRESO MENSUAL" con 650 registros, en contraste con la categoría "INGRESO ANUAL", que tuvo solo 57 registros. Para evitar posibles sesgos al modelar los días de morosidad, se optó por eliminar la columna correspondiente a "INGRESO ANUAL". Este enfoque buscó centrarse en las categorías más representativas y significativas, mejorando la precisión del modelo al capturar patrones relevantes y simplificando el conjunto de datos para una mejor generalización.



**Fig. 4.13. Número de créditos por FREC\_INGRESOS.**

Nota: Distribución del número de créditos por territorios.

En el análisis de la columna "TIPO", se observó que las categorías "REFINANCIADO" y "REESTRUCTURADO", con 6 y 2 créditos respectivamente, presentaron los valores más bajos. Además, estas categorías representan a aquellos socios que no cumplieron inicialmente con sus obligaciones financieras y tuvieron que refinanciarse. Por lo tanto, para evitar posibles complicaciones en el desarrollo del modelo, se decidió eliminar estas categorías.

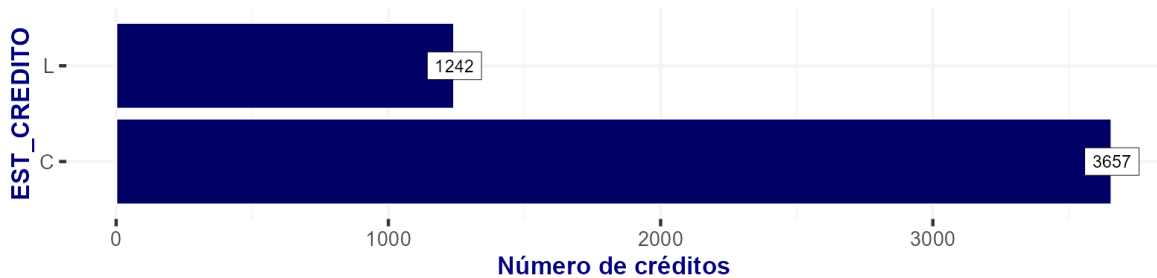


**Fig. 4.14. Número de créditos por TIPO.**

Nota: Distribución del número de créditos por tipo.



Para mejorar la predicción de los días de morosidad, se enfocó el estudio en aquellos créditos que ya han sido cancelados por sus solicitantes. En la Fig. 4.15, se destaca que los créditos cancelados (C), con un total de 3657, representan mayoritariamente a aquellos socios que han cumplido con sus obligaciones crediticias. En consecuencia, se decidió trabajar únicamente con la categoría "C" de la columna "EST\_CREDITO" para centrar el análisis en los casos donde los socios han completado el proceso crediticio.



**Fig. 4.15. Número de créditos por EST\_CREDITO.**

Nota: Distribución del número de créditos por estado de crédito.

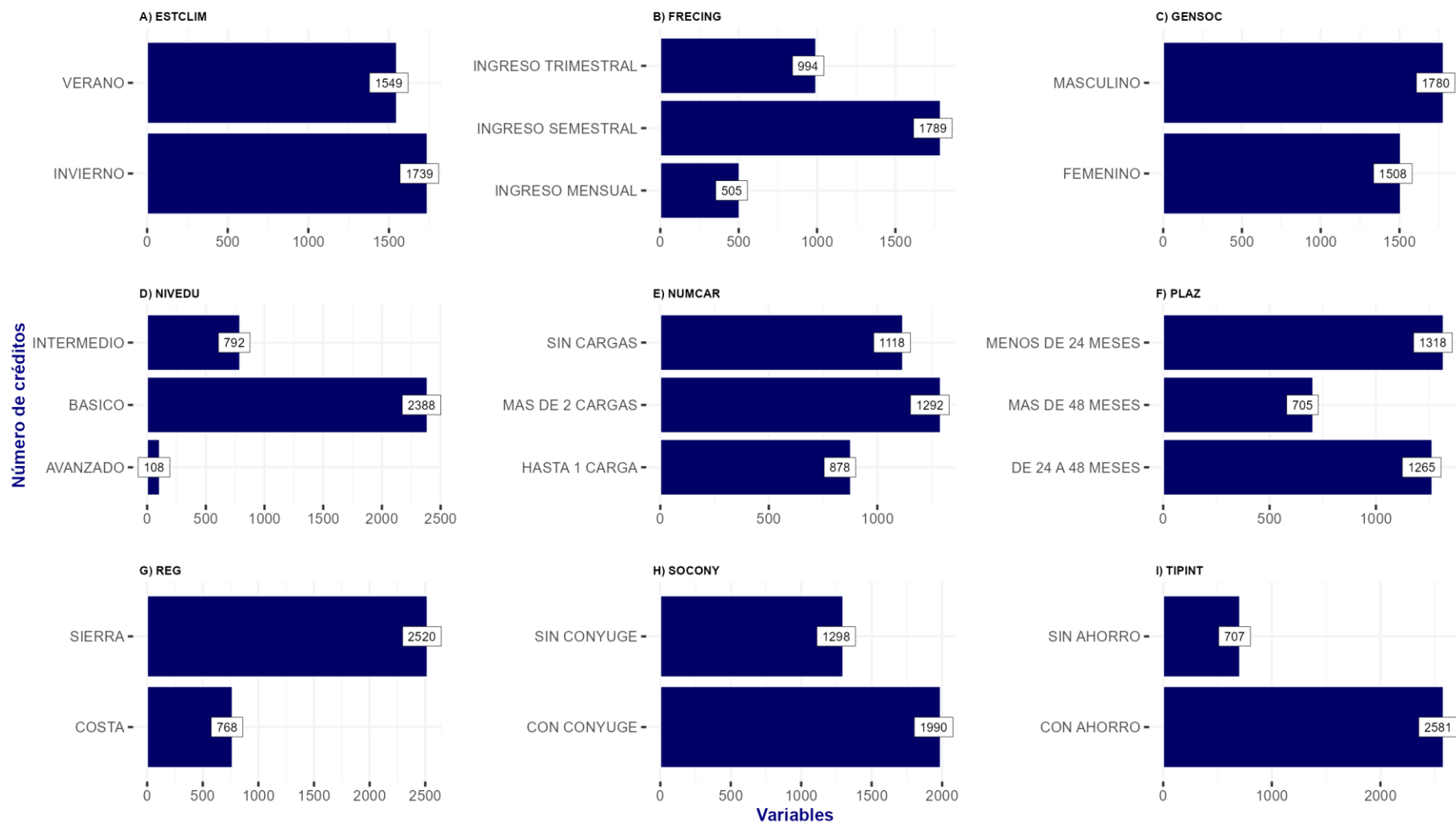
### **4.3. Descripción de cada una de las variables asociadas con la morosidad y el cumplimiento de pago, mediante técnicas estadísticas descriptivas**

#### **4.3.1. Variables categóricas**

En la Fig. 4.16 se observaron las características de los socios en el conjunto de datos analizado. En relación al género, más del 54% eran masculinos, mientras que alrededor del 45.9% eran femeninos. En cuanto al nivel educativo, la mayoría tenía educación básica (72.6%), seguido de aquellos con educación intermedia (24.1%).

En relación al estado civil, aproximadamente el 60.5% de los socios tenían cónyuge, mientras que el 39.5% no lo tenían. En términos del número de cargas, la mayoría tenía más de 2 cargas (39.3%) o ninguna carga (34.0%). En cuanto al plazo de crédito, la distribución estaba equilibrada, con una proporción significativa entre 24 y 48 meses.

En relación al tipo de interés, la mayoría de los socios tenían créditos con ahorro (78.5%). Al analizar la región geográfica, la mayoría pertenecía a la región Sierra (76.6%). En cuanto a la frecuencia de pago, la opción más común era realizar pagos con ingresos semestrales (54.4%). Finalmente, en cuanto a la estación climática, la distribución era bastante equitativa entre invierno (52.9%) y verano (47.1%).



**Fig. 4.16. Distribución de variables categóricas.**

Nota: Visualización de la distribución de las variables categóricas. a) ESTCLIM = Estación climática. b) FRECIING = Frecuencia de ingresos. c) GENSOC = Genero del socio. d) NIVEDU = Nivel de educación. e) NUNCAR = Numero de cargas. f) PLAZ = Plazo. g) REG = Región. h) SOCONY = Sociedad conyugal. i) TIPINT = Tipo de interés.

### 4.3.2. Variables cuantitativas

La edad promedio de los socios (“EDSOC”) en el estudio fue de aproximadamente 46.58 años, con una variabilidad moderada representada por una desviación estándar de 15.16. La distribución de edades exhibió una ligera asimetría positiva (0.45), indicando una concentración ligeramente mayor de socios en edades superiores al promedio. La prueba de normalidad de Shapiro-Wilk, aunque significativa ( $p < 0.001$ ), sugiere una distribución no muy alejada de la normalidad.

El ingreso mensual (“ING”) promedio de los socios fue de alrededor de 1058.32 dólares, con una variabilidad notable, como lo indica la alta desviación estándar de 919.7. La asimetría positiva (2.38) indica una distribución sesgada hacia ingresos más altos. La prueba de normalidad de Shapiro-Wilk mostró que la distribución no sigue una normalidad estricta ( $p < 0.001$ ).

En cuanto a los gastos mensuales (“GAST”), la media fue de 844.4 dólares, con una variabilidad significativa representada por una desviación estándar de 778.93. La distribución exhibió una asimetría positiva (2.49), indicando una concentración de gastos hacia valores más altos. La prueba de normalidad de Shapiro-Wilk mostró que la variable no sigue una distribución normal ( $p < 0.001$ ).

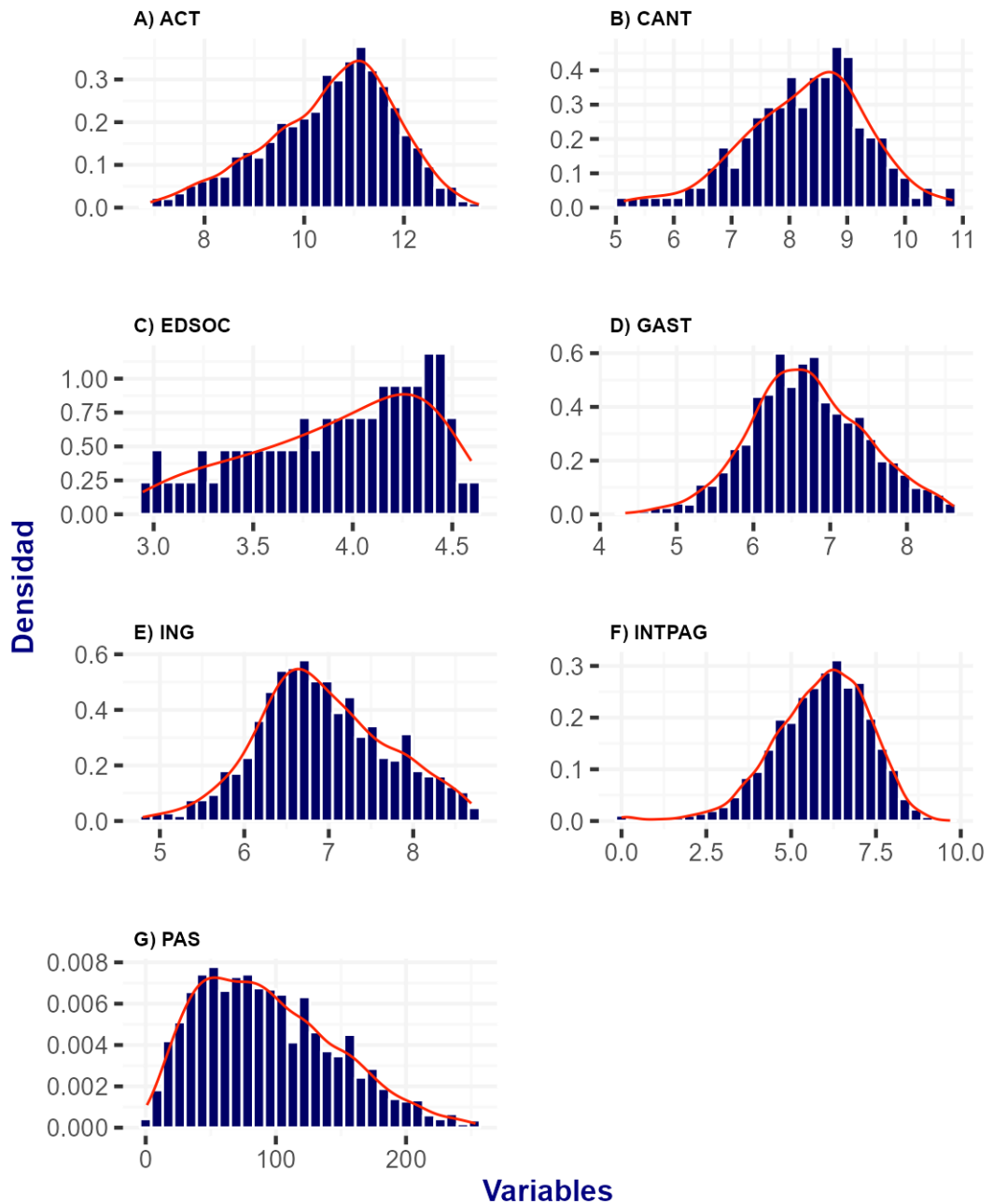
El patrimonio (“PAT”) promedio de los socios fue de alrededor de 53961.48 dólares, con una variabilidad considerable representada por una desviación estándar de 72486.41. La distribución presentó una fuerte asimetría positiva (3.35), indicando una concentración de patrimonio hacia valores más altos. La prueba de normalidad de Shapiro-Wilk indicó que la variable no sigue una distribución normal ( $p < 0.001$ ).

La cantidad (“CANT”) promedio de monto solicitado por los socios fue de 4835.05, con una variabilidad significativa (desviación estándar de 4438.84). La distribución mostró una asimetría positiva (2.83), sugiriendo una concentración de operaciones hacia valores más altos. La prueba de normalidad de Shapiro-Wilk reveló que la variable no sigue una distribución normal ( $p < 0.001$ ).

En cuanto a los intereses pagados (“INTPAG”), la media fue de 773.07 dólares, con una variabilidad notable (desviación estándar de 1075.19). La distribución exhibió una fuerte asimetría positiva (3.98), indicando una concentración de pagos de intereses hacia valores más altos. La prueba de normalidad de Shapiro-Wilk indicó que la variable no sigue una distribución normal ( $p < 0.001$ ).

La media de los activos (“ACT”) de los socios fue de alrededor de 62699.77 dólares, con una variabilidad significativa representada por una desviación estándar de 1305.13. La distribución presentó una asimetría positiva (3.20),

indicando una concentración de activos hacia valores más altos. La prueba de normalidad de Shapiro-Wilk mostró que la variable no sigue una distribución normal ( $p < 0.001$ ).



**Fig. 4.17. Histograma de columnas financieras con transformación Box-Cox.**

Nota: Visualización de la distribución de las variables numéricas. A) ACT = Activos. B) CANT = Cantidad. C) EDSOC = Edad del socio. d) GAST = Gasto. e) ING = Ingresos. F) INTPAG = Intereses pagados. G) PAS = Pasivo.

**Tabla 4.7.**  
**Estadística descriptiva para variables cuantitativas.**

Estadísticos	Variables						
	EDSOC	ING	GAST	PAT	CANT	INTPAG	ACT
N	3288	3288	3288	3288	3288	3288	3288
Perdidos	0	0	0	0	0	0	0
Media	46.58	1058.32	844.4	53961.48	4835.05	773.07	62699.77
Error estándar media	0.26	16.04	13.58	1264.13	77.41	18.75	1305.13
IC 95% de la media límite inferior*	46.06	1026.87	817.76	51482.93	4683.27	736.31	60140.82
IC 95% de la media límite superior*	47.1	1089.77	871.03	56440.04	4986.82	809.84	65258.72
Mediana	45	750	587.3	31629.57	3500	409.77	40000.00
Moda	34	600	300	0	3000	123.26	5000
Desviación estándar	15.16	919.7	778.93	72486.41	4438.84	1075.19	74837.56
Varianza	229.7	845855.31	606724.35	5254280003	19703316.91	1156030.81	5600660092.70
Mínimo	19	120	75	0	170	0	1005.00
Máximo	99	5929.6	5496.07	720292.75	50000	16112.06	733837.60
Asimetría	0.45	2.38	2.49	3.35	2.83	3.98	3.20
Error estándar asimetría	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Curtosis	-0.54	6.49	7.52	16.66	15.61	28.33	15.60
Error estándar curtosis	0.09	0.09	0.09	0.09	0.09	0.09	0.09
W de Shapiro-Wilk	0.97	0.74	0.73	0.67	0.76	0.64	0.70
Valor p de Shapiro-Wilk**	< .001	< .001	< .001	< .001	< .001	< .001	< .001
25 percentil	34	500	370.34	9810.31	2000	150.74	16000.00
50 percentil	45	750	587.3	31629.57	3500	409.77	40000.00
75 percentil	57	1200	990	68850	6000	979.74	79625.00

Nota: Estadística descriptiva, para los variables cuantitativas. ACT = Activos. CANT = Cantidad. EDSOC = Edad socio. GAST = Gastos. ING = Ingresos. INTPAG = Intereses pagados. PAS = Pasivos. PAT = Patrimonio.

\*El IC de la media supone que las medias muestrales siguen una distribución t con N - 1 grados de libertad.

\*\*Prueba de Shapiro-Wilk, es significativa al 0.01 (bilateral):

$$H_0 = \sim N(\mu, \sigma^2); H_1 = \neq N(\mu, \sigma^2)$$

### 4.3.3. Variable respuesta

#### 4.3.3.1. Análisis descriptivo

La variable respuesta DMORA, que representaba los días de morosidad, se caracterizó por tener un total de 3288 observaciones. La media de días de morosidad fue de 36.6, con un error estándar de la media de 1.13. El intervalo de confianza del 95% para la media varió entre 34.39 y 38.8.

La mediana fue de 7, indicando que el 50% de las observaciones tenían 7 días o menos de morosidad. La moda, fue 0, sugiriendo que había una concentración de observaciones con ningún día de morosidad.

La desviación estándar de 64.51 indicó la dispersión de los datos alrededor de la media, y la varianza fue de 4161.53. La asimetría de la distribución fue 2.51, sugiriendo una distribución sesgada hacia la izquierda. La curtosis fue de 6.34, mostrando colas más pesadas de lo esperado para una distribución normal.

El valor p del test de Shapiro-Wilk fue menor que 0.001, indicando que los datos no seguían una distribución normal. Los percentiles mostraron que el 25% de las observaciones tenían 0 días o menos de morosidad, el 50% tenían 7 días o menos, y el 75% tenían 40 días o menos de morosidad.

**Tabla 4.8.**  
**Estadísticas descriptiva de la variable días de morosidad.**

Estadísticos	Valor
N	3288
Media	36.6
Error estándar media	1.13
IC 95% de la media límite inferior*	34.39
IC 95% de la media límite superior*	38.8
Mediana	7
Moda	0
Desviación estándar	64.51
Varianza	4161.53
Mínimo	0
Máximo	360
Asimetría	2.51
Curtosis	6.34
W de Shapiro-Wilk	0.62
Valor p de Shapiro-Wilk**	< .001
25 percentil	0
50 percentil	7
75 percentil	40

Nota: Estadística descriptiva, para los días de morosidad (DMORA).

\*El IC de la media supone que las medias muestrales siguen una distribución t con N - 1 grados de libertad.

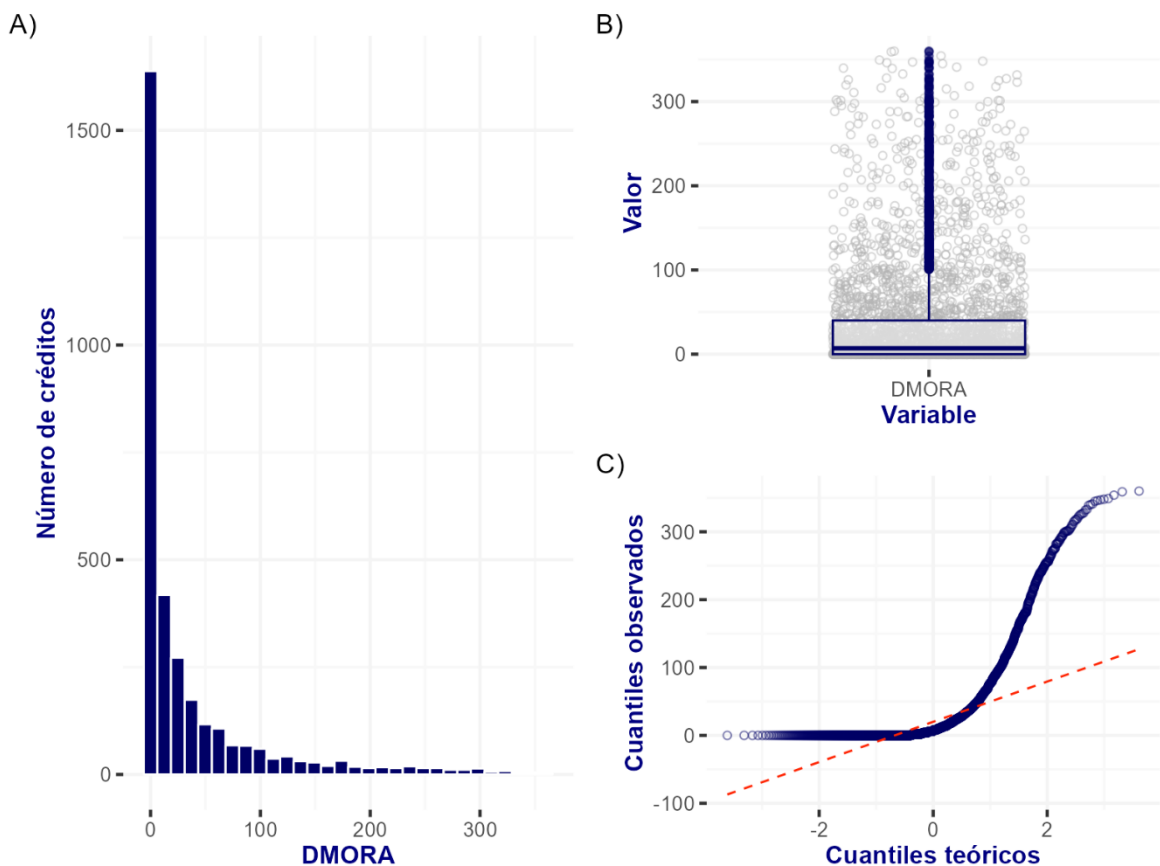
\*\*Prueba de Shapiro-Wilk, es significativa al 0.01 (bilateral):

$H_0 = \sim N(\mu, \sigma^2); H_1 = \neq N(\mu, \sigma^2)$

En el histograma de DMORA, se observó que la mayoría de los datos se concentraba hacia el extremo inferior del eje X, muy cercanos a cero. Este patrón indicaría una mayor concentración de valores bajos de DMORA, sugiriendo que la mayoría de las observaciones tenían pocos días de morosidad.

El gráfico de boxplot proporcionó información sobre la dispersión y concentración de los valores de DMORA. Se observó que la mediana y la mayor parte de los datos estaban dentro del rango intercuartil (IQR), aunque también se identificaron puntos fuera de los bigotes, considerados como valores atípicos, lo que indicó la presencia de observaciones inusuales en el conjunto de datos.

En el gráfico QQ, se compararon los cuantiles observados de la variable DMORA con los cuantiles teóricos de una distribución normal. Se notó una desviación significativa de la línea de referencia para cuantiles teóricos de una distribución normal, especialmente en los extremos. Esto demostró que DMORA no seguía una distribución normal y podía presentar un sesgo hacia valores más bajos y altos de morosidad.



**Fig. 4.18. Visualización de la variable respuesta DMORA.**

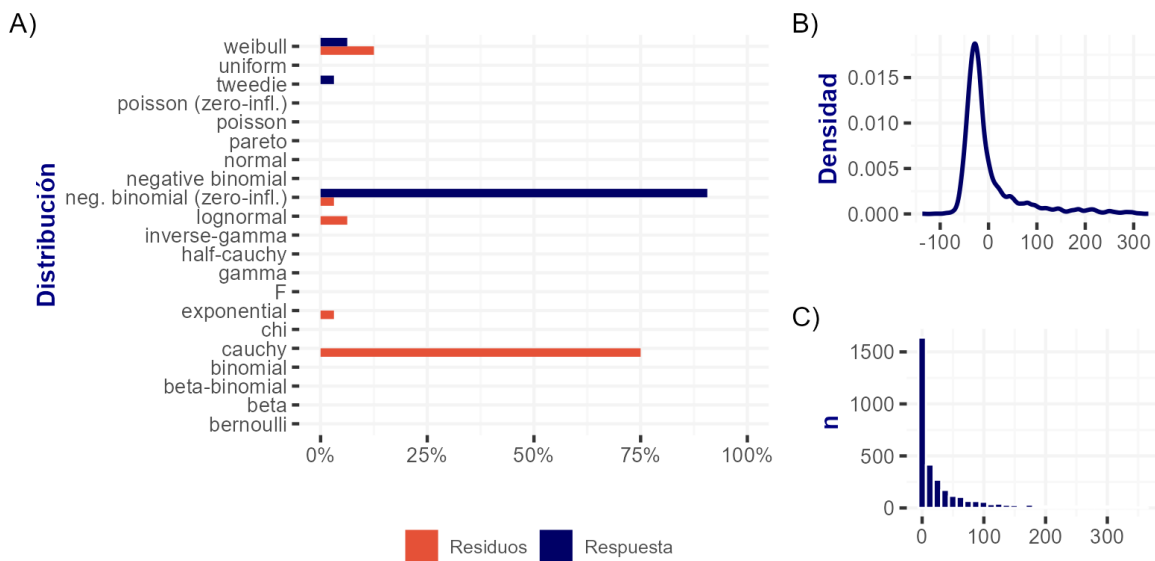
Nota: Exploración Visual de la Distribución de DMORA. A) Histograma de DMORA, B) Boxplot para DMORA, C) Gráfico de QQ para DMORA.

En la Fig. 4.19, se observó que la variable DMORA fue clasificada como una distribución binomial negativa inflada de ceros, relacionada con el tiempo que tomó para que una persona entrara en morosidad. La inflación de ceros indicó una prevalencia significativa de casos con respuesta cero, señalando una alta probabilidad de no ocurrencia del evento en esos casos específicos.

La regresión binomial negativa inflada de ceros se utiliza en situaciones con datos discretos que representan la frecuencia de eventos, especialmente cuando se observa una cantidad considerable de casos con respuesta cero. Este modelo resultó útil para abordar la sobredispersión, donde la variabilidad de los datos es mayor de lo esperado en un modelo binomial estándar [63].

Este enfoque además puede entender como un modelo mixto de dos componentes, donde uno de ellos sigue una distribución declinada por ceros, y el otro se ajusta mediante un modelo de regresión estándar para datos de recuento.

Adicionalmente, el análisis reveló que los residuos de la variable DMORA presentaban una distribución Cauchy. Esta distribución, conocida por su cola más pesada y sensibilidad a valores atípicos, sugiere la posible presencia de observaciones inusuales en la variable DMORA, información relevante al interpretar los resultados del modelo.



**Fig. 4.19. Familias de distribución de la morosidad (DMORA).**

Nota: Clasificación de la distribución de la familia de la variable respuesta morosidad (DMORA), utilizando un modelo regresión (1m) por medio de una técnica de bosque aleatorio interno. A) Distribución prevista de residuales y respuesta. B) Densidad de residuos. C) Distribución de respuesta.



## 4.4. Construcción modelos predictivos para el cumplimiento de pago, mediante técnicas estadísticas y de minería de datos.

### 4.4.1. Modelo de regresión Poisson

La Tabla 4.9 presenta información sobre el modelo de regresión de Poisson, utilizado para la predicción de los días de mora en créditos productivos con acompañamiento técnico. Según la Tabla 4.9, el modelo fue desarrollado usando función de R base, `glm()`, conocida por ser útil en la implementación de modelos lineales generalizados. Además, se aplicó una función de enlace logarítmica (`log`) para la construcción del modelo.

Las métricas de ajuste, representadas por el AIC y BIC obtuvieron valores de 178614.7 y 178735.3 respectivamente. La desviación residual, evidenció un valor de 171102 en comparación con la desviación nula de 188375, indicando una mejora significativa en la capacidad del modelo para explicar la variabilidad en los días de mora al incluir variables independientes.

En los grados de libertad, se observó una diferencia de 20 grados entre los asociados con el modelo nulo (2301) y los residuales del modelo (2281), resaltando la contribución significativa de las variables predictoras al ajuste general del modelo. El bajo número de iteraciones Fisher Scoring (6) indicaron una rápida convergencia del modelo durante el proceso de ajuste, señalando la estabilidad y fiabilidad del mismo.

**Tabla 4.9.**  
**Resumen del modelo de regresión Poisson.**

Información	Valor	Comentario
Tipo de modelo	Poisson	Modelo para datos de conteo
Método	<code>glm</code>	<code>glm(DMORA~., family = poisson(link = "log"), data = datos_ entrenamiento)</code>
Función de enlace	<code>log</code>	Coefficientes del modelo en escala <code>log()</code>
Distribución	Poisson	Modelo para datos de conteo
AIC	178614.7	Métrica del ajuste del modelo
BIC	178735.3	Métrica del ajuste del modelo
Null deviance	188375	Métrica de ajuste nulo
Residual deviance	171102	Métrica de ajuste incluidas variables predictoras
gL, nulos	2301	gL asociados con el modelo nulo
gL, residuales	2281	gL residuales del modelo
Iteraciones Fisher Scoring	6	

Nota: Información del modelo de regresión. Destaque las métricas AIC y BIC, donde valores más bajos indican un mejor ajuste. AIC = Criterio de Información de Akaike, BIC = Criterio de Información Bayesiano, Null deviance = Desviación nula, Residual deviance = Desviación residual. gL: null = grados de libertad nulos, gL: residual = grados de libertad residuales.

En el análisis de los coeficientes estimadores ( $\beta$ ) y la IRR ( $e^\beta$ ) en el modelo de regresión de Poisson, se destacó que los predictores con valores de IRR mayores a 1 estaban asociados con un aumento en la tasa de incidencia de morosidad. Por otro lado, aquellos con IRR menores a 1 indicaban una disminución en la tasa de incidencia.

Todas las variables analizadas en el modelo de regresión de Poisson presentaron niveles significativos de confianza, superando el umbral del 95%. Esta alta significancia estadística fortaleció la validez de las conclusiones derivadas del análisis, respaldando la influencia de cada variable en la predicción de la tasa de incidencia de días de morosidad.

El intercepto del modelo obtuvo un coeficiente estimado de 2.606; por lo tanto, manteniendo constantes las demás variables, se esperaba que la tasa de incidencia de días de morosidad fuera aproximadamente 13.54 veces mayor.

La variable EDSOC reveló un coeficiente estimado de 0.2302, indicando un aumento del 26% en la tasa de incidencia de días de mora con el incremento de la edad del socio. La variable ING mostró un coeficiente negativo de -0.1878, indicando que un aumento en los ingresos se asociaba con una disminución del 17.7% en la tasa de incidencia de días de mora.

El coeficiente positivo de GAST fue de 0.1253, sugiriendo que un aumento en los gastos estaba relacionado con un incremento del 12.53% en la tasa de incidencia de días de mora. Por otro lado, la variable ACT presentó un coeficiente negativo de -0.224, señalando una disminución del 22.4% en la tasa de incidencia de días de mora con un aumento en los activos.

La variable CANT mostró un coeficiente de 0.2189, indicando un aumento del 24% en la tasa de incidencia de días de mora. De manera similar, INTPAG presentó un coeficiente positivo de 0.1879, sugiriendo un aumento del 18.79% en la tasa de incidencia de días de mora con un incremento en los intereses pagados.

En cuanto al género, se observó que las personas categorizadas como femeninas (GÉNERO [2]) tenían una tasa de incidencia de días de morosidad aproximadamente un 7% mayor en comparación con la categoría de referencia.

El nivel educativo también influyó, mostrando que aquellos con educación primaria (NIVEL EDUCATIVO [2]) tenían una tasa de incidencia un 22% mayor, mientras que aquellos con educación secundaria (NIVEL EDUCATIVO [3]) presentaban un aumento del 6% en la tasa de incidencia.

La presencia de sociedad conyugal (SOCONY [1]) estuvo asociada con una disminución del 20% en la tasa de incidencia de morosidad en comparación con aquellos sin sociedad conyugal.

En relación con el número de cargas familiares (NUMCAR), tener dos cargas (NUMCAR [2]) se asoció con una disminución del 5%, mientras que tener más de tres cargas (NUMCAR [3]) aumentó la tasa de incidencia en un 6%.

El plazo del crédito (PLAZ) también influyó, mostrando que tanto un plazo mediano (PLAZ [2]) como un plazo largo (PLAZ [3]) estuvieron asociados con disminuciones del 3% y 4%, respectivamente, en la tasa de incidencia de la morosidad en comparación con un plazo corto.

El tipo de interés (TIPINT [1]) indicó que aquellos con ahorro tenían una tasa de incidencia un 5% menor en comparación con aquellos sin ahorro. En cuanto a la región geográfica (REG), la sierra (REG [2]) presentó una disminución del 8% en la tasa de incidencia de la morosidad en comparación con la costa.

La frecuencia de ingresos agropecuarios (FRECING) mostró que una frecuencia trimestral (FRECING [2]) se asoció con un aumento del 5%, mientras que una frecuencia semestral (FRECING [3]) estuvo relacionada con una disminución del 14.81%.

La estación climática (ESTCLIM) reveló que el verano (ESTCLIM [2]) se asoció con un aumento del 15% en la tasa de incidencia de la morosidad en comparación con el invierno.

El modelo, en su conjunto, demostró una alta capacidad para explicar la variabilidad en la tasa de días de mora, como se refleja en las métricas de ajuste como la devianza y el AIC que respaldaron la idoneidad del modelo para explicar los patrones observados.

**Tabla 4.10.**  
**Parámetros del modelo de regresión Poisson.**

Predictores	$\beta$	IRR ( $e^\beta$ )	IC	F	p
(Intercepto)	2.606	13.54	11.66 – 15.73	34.07	<0.001 ***
EDSOC	0.2302	1.26	1.22 – 1.29	16.48	<0.001 ***
ING	-0.1878	0.83	0.81 – 0.84	-21.52	<0.001 ***
GAST	0.1253	1.13	1.11 – 1.15	13.64	<0.001 ***
ACT	-0.224	0.8	0.79 – 0.81	-52.94	<0.001 ***
PAS	0.001461	1.0	1.00 – 1.00	18.25	<0.001 ***
CANT	0.2189	1.24	1.22 – 1.27	25.02	<0.001 ***
INTPAG	0.1879	1.21	1.19 – 1.22	33.06	<0.001 ***
GENSOC [2]	0.06455	1.07	1.05 – 1.08	8.55	<0.001 ***
NIVEDU [2]	0.1964	1.22	1.20 – 1.24	24.2	<0.001 ***
NIVEDU [3]	0.05964	1.06	1.02 – 1.10	3.19	0.001 **
SOCONY [1]	-0.2206	0.8	0.79 – 0.82	-26.22	<0.001 ***
NUMCAR [2]	-0.04918	0.95	0.93 – 0.97	-5.26	<0.001 ***
NUMCAR [3]	0.0565	1.06	1.04 – 1.08	6.39	<0.001 ***
PLAZ [2]	-0.0343	0.97	0.95 – 0.99	-3.46	0.001 ***
PLAZ [3]	-0.04114	0.96	0.94 – 0.98	-3.14	0.002 **
TIPINT [1]	-0.04859	0.95	0.93 – 0.97	-5.1	<0.001 ***
REG [2]	-0.07892	0.92	0.90 – 0.94	-7.14	<0.001 ***
FRECING [2]	0.0459	1.05	1.02 – 1.07	4.15	<0.001 ***
FRECING [3]	-0.1481	0.86	0.85 – 0.88	-14.66	<0.001 ***
ESTCLIM [2]	0.1399	1.15	1.13 – 1.17	20.29	<0.001 ***
Observaciones	2302				
Deviance	171102.336				
AIC	178614.721				
log-Likelihood	-89286.361				

Nota: Resumen de estadísticas del modelo de regresión para validación.  $\beta$  = Coeficientes estimados, IRR = Incidence Rate Ratio (*Razón de Tasas de Incidencia*), IC = Intervalo de confianza, F = Estadística F de Fisher, p = Valor p (probabilidad). Niveles de Significancia: 0 \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05 \* 0.1 ' ' 1.

#### 4.4.2. Modelo de regresión binomial negativa

El modelo de regresión binomial negativo se implementó utilizando la función `glm.nb()` de la librería MASS, con una función de enlace logarítmica (log) para los coeficientes del modelo, indicando que estos operan en la escala logarítmica (Tabla 4.12).

Al evaluar el ajuste del modelo, se observaron valores bajos tanto para el AIC (17830.32) como para el BIC (17956.63), indicando un buen ajuste. La varianza nula (null deviance) fue de 2570.7, y la varianza residual (residual deviance) fue 2486.2. La disminución en la varianza al incluir variables predictoras señaló una mejora en la capacidad del modelo para explicar la variabilidad en los días de morosidad (DMORA). Estos resultados se respaldaron con los grados de libertad asociados a la varianza nula y residual, que fueron 2301 y 2281, respectivamente.

El parámetro de ajuste (theta) tuvo un valor de 0.22897, y el bajo error estándar de 0.00705 sugiere una alta precisión en los parámetros estimados. La métrica de

ajuste del modelo, representada por el doble del logaritmo de la verosimilitud (-17784.648), demuestra la capacidad del modelo para describir los días de morosidad.

**Tabla 4.11.**  
**Resumen del modelo de regresión binomial negativa.**

Información	Valor	Comentario
Tipo de modelo	Binomial negativa	Modelo para datos de conteo
Librería	MASS	
Método	glm.nb	glm.nb(DMORA~., data = datos_train)
Función de enlace	log	Coefficientes del modelo en escala log()
Distribución	Binomial negativa	Modelo para datos de conteo
AIC	17830.32	Métrica de ajuste del modelo
BIC	17956.63	Métrica de ajuste del modelo
Null deviance	2570.7	Métrica de ajuste nulo
Residual deviance	2486.2	Métrica de ajuste incluidas variables predictoras
gL; null	2301	gL asociados con el modelo nulo
gL, residual	2281	gL residuales del modelo
Theta	0.22897	Parámetro de ajuste del modelo
Error estándar	0.00705	Medida de estimación de los parámetros
2 x log-likelihood	-17784.648	Métrica de ajuste del modelo

Nota: Información del modelo de regresión. Destaque las métricas AIC y BIC, donde valores más bajos indican un mejor ajuste. AIC = Criterio de Información de Akaike, BIC = Criterio de Información Bayesiano, Null deviance = Desviación nula, Residual deviance = Desviación residual. gL: null = grados de libertad nulos, gL: residual = grados de libertad residuales.

El análisis de los coeficientes estimados del modelo de regresión binomial negativo reveló información clave sobre la asociación de diversos predictores con la tasa de incidencia de morosidad (Tabla 4.12). El intercepto, con un coeficiente estimado ( $\beta$ ) de 1.6203, indicó que, manteniendo constantes las demás variables, la tasa de incidencia era aproximadamente 5.05 veces mayor.

La variable EDSOC (edad del socio) obtuvo un coeficiente estimado de 0.4367, sugiriendo que un incremento en la edad del socio se asociaba con un aumento del 55% en la tasa de incidencia de morosidad. Además, la variable GAST (gastos) mostró un coeficiente positivo de 0.1951, indicando que un aumento en los gastos estaba relacionado con un incremento del 22% en la tasa de incidencia.

Los activos (ACT) presentaron un coeficiente negativo significativo de -0.2784, señalando una disminución del 24% en la tasa de incidencia con su aumento. Asimismo, la variable CANT (cantidad o monto del crédito) tenía un coeficiente positivo de 0.221, indicando un aumento del 25% en la tasa de incidencia.

En cuanto a los intereses pagados (INTPAG), un coeficiente estimado de 0.1665 sugería que un aumento en esta variable estaba asociado con un incremento del 18.79% en la tasa de incidencia. Por último, la presencia de sociedad conyugal (SOCONY [1]) reveló un coeficiente negativo de -0.2605, indicando una

disminución del 23% en la tasa de incidencia de morosidad en comparación con aquellos sin sociedad conyugal.

En relación con las métricas del modelo, se analizaron 2302 observaciones, y el R<sup>2</sup> Nagelkerke reveló una proporción de variabilidad explicada relativamente baja, alcanzando 0.066. La deviance, con un valor de 2487.112, indicó cómo el modelo se ajusta a los datos, y métricas como el AIC (17737.86) y log verosimilitud (-8846.93) proporcionaron información sobre la bondad del ajuste, buscando minimizar la complejidad del modelo y maximizar la verosimilitud para una interpretación más precisa de la tasa de incidencia de morosidad.

**Tabla 4.12.**  
**Parámetros del modelo de regresión binomial negativa.**

Predictores	$\beta$	IRR ( $e^{\beta}$ )	IC	F	p
(Intercept)	1.6203	5.05	0.78 – 32.65	1.73	<b>0.084 .</b>
EDSOC	0.4367	1.55	1.10 – 2.17	2.53	<b>0.011 *</b>
ING	-0.1458	0.86	0.69 – 1.08	-1.34	0.181
GAST	0.1951	1.22	0.96 – 1.54	1.7	<b>0.088 .</b>
ACT	-0.2784	0.76	0.68 – 0.85	-5.22	<b>&lt;0.001 ***</b>
PAS	0.0007	1	1.00 – 1.00	0.7	0.484
CANT	0.2210	1.25	1.03 – 1.53	2.31	<b>0.021 *</b>
INTPAG	0.1665	1.18	1.05 – 1.32	2.98	<b>0.003 **</b>
GENSOC [2]	0.0255	1.03	0.85 – 1.23	0.27	0.788
NIVEDU [2]	0.1923	1.21	0.98 – 1.51	1.79	<b>0.073 .</b>
NIVEDU [3]	0.0700	1.07	0.66 – 1.86	0.27	0.786
SOCONY [1]	-0.2605	0.77	0.63 – 0.95	-2.48	<b>0.013 *</b>
NUMCAR [2]	-0.0076	0.99	0.79 – 1.25	-0.07	0.948
NUMCAR [3]	0.0249	1.03	0.83 – 1.27	0.22	0.823
PLAZ [2]	0.0820	1.09	0.85 – 1.39	0.68	0.495
PLAZ [3]	-0.0052	0.99	0.70 – 1.41	-0.03	0.975
TIPINT [1]	-0.1068	0.9	0.70 – 1.14	-0.88	0.381
REG [2]	0.0129	1.01	0.76 – 1.34	0.09	0.929
FRECING [2]	0.2390	1.27	0.95 – 1.68	1.64	0.101
FRECING [3]	-0.0132	0.99	0.76 – 1.27	-0.1	0.921
ESTCLIM [2]	0.1485	1.16	0.98 – 1.38	1.71	<b>0.088 .</b>
Observaciones	2302				
Deviance	2487.112				
AIC	17737.86				
log-Likelihood	-8846.93				

Nota: Resumen de estadísticas del modelo de regresión para validación.  $\beta$  = Coeficientes estimados, IRR = Incidence Rate Ratio (*Razón de Tasas de Incidencia*), IC = Intervalo de confianza, F = Estadística F de Fisher, p = Valor p (probabilidad). Niveles de Significancia: 0 \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05 . 0.1 ' 1.

#### 4.4.3. Modelo de regresión Poisson inflada de ceros

En el modelo de regresión Poisson inflada de ceros, se empleó la función `zeroinfl()` de la librería `pscl` para modelar datos de recuentos con exceso de

ceros. La función de enlace para el conteo fue Log, indicando que los coeficientes del modelo operan en la escala logarítmica. Por otro lado, la función de enlace para la inflación de ceros fue Logit, que se utiliza para modelar las probabilidades de tener ceros o no en los datos.

En relación con la distribución, se utilizó Poisson para los datos de conteo y Binomial para la inflación de ceros, lo cual es característico de modelos inflados de ceros. El AIC y BIC del modelo fueron 115201.3 y 115442.5 respectivamente, métricas que evalúan la bondad de ajuste del modelo, siendo preferibles valores más bajos.

La log-verosimilitud del modelo fue -57560, y los grados de libertad asociados con la log-verosimilitud fueron 42. Estos indicadores son relevantes para evaluar la capacidad del modelo para describir los datos observados.

En el proceso de optimización BFGS, se realizaron 1 iteración, señalando la eficiencia del algoritmo de optimización en la convergencia del modelo durante el ajuste.

**Tabla 4.13.**  
**Resumen del modelo de regresión Poisson inflada de ceros.**

Información	Valor	Comentario
Tipo de modelo	Poisson inflada de ceros	Modelo para datos de recuentos con exceso de ceros
Librería	pscl	
Método	zeroinfl	zeroinfl(DMORA~.,link = "logit", dist = "poisson", data=datos_train)
Función de enlace (conteo)	Log	Coefficientes del modelo en escala log()
Función de enlace (inflación cero)	Logit	Log de probabilidades, y=1 sobre y=0
Distribución (conteo)	Poisson	Modelo para datos de conteo
Distribución (inflación cero)	Binomial	Modelo para datos binario
AIC	115201.3	Métrica del ajuste del modelo
BIC	115442.5	Métrica del ajuste del modelo
Log-likelihood	-57560.00	Métrica del ajuste del modelo
gL: Log-likelihood	42	gL asociados con la log-verosimilitud
Iteración, optimización BFGS	1	Numero de iteraciones

Nota: Información del modelo de regresión. Destaque las métricas AIC y BIC, donde valores más bajos indican un mejor ajuste. AIC = Criterio de Información de Akaike, BIC = Criterio de Información Bayesiano, Null deviance = Desviación nula, Residual deviance = Desviación residual. gL: null = grados de libertad nulos, gL: residual = grados de libertad residuales, gL: (Log-likelihood) = grados de libertad asociados con la log-verosimilitud. Iteración, BFGS = Número de iteración del algoritmo de optimización Broyden-Fletcher-Goldfarb-Shanno.

En el análisis del modelo de regresión Poisson inflado con ceros, realizado en la parte de conteo, se examinaron los coeficientes estimados de varios predictores que demostraron ser significativos para explicar los días de morosidad. El intercepto, con un coeficiente estimado ( $\beta$ ) de 2.748, indicó que, manteniendo

constantes las demás variables, la tasa de incidencia es aproximadamente 15.61 veces mayor.

La edad del socio (EDSOC) presentó un aumento del 25% en la tasa de incidencia de morosidad, con un coeficiente estimado de 0.2218. Por otro lado, los ingresos (ING) demostraron un coeficiente negativo significativo de -0.0599, indicando una disminución del 6% en la tasa de incidencia con su aumento.

Los activos (ACT) registraron un coeficiente negativo significativo de -0.1332 en el modelo de regresión, señalando una disminución del 12% en la tasa de incidencia de morosidad por cada unidad adicional en la variable de activos.

En relación a los pasivos (PAS), se observó un ligero aumento del 0.04% en la tasa de incidencia, reflejado por el coeficiente positivo de 0.0004. Respecto a la cantidad (CANT), su coeficiente positivo de 0.2024 se asoció con un incremento significativo del 22% en la tasa de incidencia de morosidad, indicando que mayores montos de crédito están relacionados con un aumento proporcional en la tasa de incidencia.

Los intereses pagados (INTPAG) presentaron un coeficiente estimado de 0.0609, sugiriendo un incremento del 6.39% en la tasa de incidencia con cada unidad adicional en esta variable. El género del socio (GENSOC [2]) mostró un coeficiente negativo de -0.0247, indicando una disminución del 2% en la tasa de incidencia para aquellos con dicho género en comparación con los demás.

En relación al nivel educativo, tanto NIVEDU [2] como NIVEDU [3] presentaron coeficientes positivos, demostraron aumentos en la tasa de incidencia asociados con niveles educativos más altos. La sociedad conyugal (SOCONY [1]) reveló un coeficiente negativo de -0.1837, señalando una disminución del 18% en la tasa de incidencia para aquellos con sociedad conyugal en comparación con los individuos sin este estado civil.

El número de cargas (NUMCAR [3]) exhibió un coeficiente positivo de 0.0645, indicando un aumento del 6.62% en la tasa de incidencia de morosidad con cada unidad adicional en esta variable. En cuanto al plazo del préstamo (PLAZ [2] y PLAZ [3]), ambos presentaron coeficientes positivos, lo que demuestra aumentos en la tasa de incidencia asociados con plazos más largos.

El tipo de interés (TIPINT [1]) demostró un coeficiente negativo de -0.0569, sugiriendo una disminución del 5.69% en la tasa de incidencia asociada con dicho tipo de interés. Respecto a la frecuencia de pago (FRECING [2] y FRECING [3]), ambas variables presentaron coeficientes positivos, indicando que mayores frecuencias de pago están asociadas con incrementos en la tasa de incidencia.



Finalmente, la estación climática (ESTCLIM [2]) demostró un coeficiente positivo de 0.0529, sugiriendo un aumento del 5.29% en la tasa de incidencia para condiciones climáticas específicas. Estos resultados proporcionan una comprensión detallada de la influencia de diversos factores en la tasa de incidencia de morosidad.

Por otro lado, en el modelo inflado a cero, se determinaron los coeficientes de predictores específicos para la probabilidad de tener ceros en la variable de respuesta, exclusivamente en las predictores que presentaron una significancia estadística igual o superior al 95%.

Para la variable EDSOC (edad del socio), se observó un coeficiente negativo de -0.4647, demostrando una disminución del 37% en la probabilidad de tener ceros a medida que aumenta la edad del socio. Respecto a GAST (gastos), el coeficiente positivo de 0.0175 indicó un incremento del 1.75% en la probabilidad de que la variable de respuesta tenga ceros con mayores niveles de gastos.

En cuanto a ACT (activos), el coeficiente positivo de 0.3126 demostró un aumento del 31.26% en la probabilidad de tener ceros en la variable de respuesta con el incremento de los activos. El predictor PAS (pasivos) presentó un coeficiente negativo de -0.0031, sugiriendo una disminución del 0.31% en la probabilidad de tener ceros con mayores niveles de pasivos. Para INTPAG (intereses pagados), el coeficiente negativo de -0.1935 indicó una disminución del 19.35% en la probabilidad de tener ceros con mayores intereses pagados.

Finalmente, SOCONY [1] (sociedad conyugal) mostró un coeficiente positivo de 0.3017, señalando un aumento del 30.17% en la probabilidad de que la variable de respuesta presente ceros en comparación con aquellos individuos sin sociedad conyugal.

En cuanto a las métricas del modelo, se analizaron 2302 observaciones. La deviance, con un valor de 107618.322, indicó cómo el modelo se ajusta a los datos. Además, se obtuvo el AIC (107618.322) y el log-Likelihood (-53767.161), ofreciendo información sobre la bondad del ajuste. Estas métricas buscaron minimizar la complejidad del modelo y maximizar la verosimilitud para una interpretación más precisa de la tasa de incidencia de morosidad.

**Tabla 4.14.**  
**Parámetros del modelo de regresión Poisson inflado con ceros.**

Predictores	$\beta$	IRR ( $e^{\beta}$ )	IC	F	p
<b>Modelo de recuento</b>					
(Intercepto)	2.7480	15.61	13.38 – 18.21	34.93	<0.001 ***
EDSOC	0.2218	1.25	1.21 – 1.28	15.86	<0.001 ***
ING	-0.0599	0.94	0.93 – 0.96	-6.75	<0.001 ***
GAST	0.0175	1.02	1.00 – 1.04	1.87	0.062 .
ACT	-0.1332	0.88	0.87 – 0.88	-30.81	<0.001 ***
PAS	0.0004	1	1.00 – 1.00	4.75	<0.001 ***
CANT	0.2024	1.22	1.20 – 1.25	22	<0.001 ***
INTPAG	0.0609	1.06	1.05 – 1.08	10.43	<0.001 ***
GENSOC [2]	-0.0247	0.98	0.96 – 0.99	-3.22	0.001 **
NIVEDU [2]	0.1904	1.21	1.19 – 1.23	23.11	<0.001 ***
NIVEDU [3]	0.0774	1.08	1.04 – 1.12	3.83	<0.001 ***
SOCONY [1]	-0.1837	0.83	0.82 – 0.85	-21.45	<0.001 ***
NUMCAR [2]	0.0182	1.02	1.00 – 1.04	1.92	0.054 .
NUMCAR [3]	0.0645	1.07	1.05 – 1.09	7.11	<0.001 ***
PLAZ [2]	0.0460	1.05	1.03 – 1.07	4.39	<0.001 ***
PLAZ [3]	0.0982	1.1	1.07 – 1.13	7.1	<0.001 ***
TIPINT [1]	-0.0569	0.94	0.93 – 0.96	-5.61	<0.001 ***
REG [2]	-0.0191	0.98	0.96 – 1.00	-1.58	0.115
FRECING [2]	0.1652	1.18	1.15 – 1.21	14.32	<0.001 ***
FRECING [3]	-0.0382	0.96	0.94 – 0.98	-3.61	<0.001 ***
ESTCLIM [2]	0.0529	1.05	1.04 – 1.07	7.53	<0.001 ***
<b>Modelo inflado a cero</b>					
(Intercepto)	0.2488	1.28	0.19 – 8.55	0.26	0.797
EDSOC	-0.4647	0.63	0.44 – 0.89	-2.6	0.009 **
ING	0.1222	1.13	0.91 – 1.41	1.08	0.28
GAST	-0.3248	0.72	0.57 – 0.91	-2.74	0.006 **
ACT	0.3126	1.37	1.22 – 1.53	5.57	<0.001 ***
PAS	-0.0031	1	0.99 – 1.00	-2.93	0.003 **
CANT	0.0319	1.03	0.85 – 1.25	0.33	0.743
INTPAG	-0.1935	0.82	0.74 – 0.92	-3.43	0.001 ***
GENSOC [2]	-0.1354	0.87	0.72 – 1.06	-1.38	0.167
NIVEDU [2]	0.0234	1.02	0.82 – 1.27	0.21	0.834
NIVEDU [3]	-0.0459	0.96	0.56 – 1.63	-0.17	0.866
SOCONY [1]	0.3017	1.35	1.09 – 1.67	2.76	0.006 **
NUMCAR [2]	-0.0505	0.95	0.75 – 1.20	-0.42	0.673
NUMCAR [3]	-0.0701	0.93	0.74 – 1.17	-0.61	0.545
PLAZ [2]	-0.1245	0.88	0.69 – 1.12	-1.01	0.313
PLAZ [3]	0.0774	1.08	0.77 – 1.51	0.45	0.651
TIPINT [1]	0.0922	1.1	0.86 – 1.40	0.73	0.466
REG [2]	-0.2319	0.79	0.59 – 1.06	-1.55	0.122
FRECING [2]	0.0648	1.07	0.79 – 1.44	0.43	0.67
FRECING [3]	0.1456	1.16	0.88 – 1.52	1.06	0.291
ESTCLIM [2]	-0.1201	0.89	0.74 – 1.06	-1.33	0.183
Observaciones	2302				
AIC	107618.322				
log-Likelihood	-53767.161				

Nota: Resumen de estadísticas del modelo de regresión para validación.  $\beta$  = Coeficientes estimados, IRR = Incidence Rate Ratio (*Razón de Tasas de Incidencia*), IC = Intervalo de confianza, F = Estadística F de Fisher, p = Valor p (probabilidad). Niveles de Significancia: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 '.' 1.

#### 4.4.4. Modelo de regresión binomial negativa inflada de ceros

En el modelo de regresión binomial negativa inflada de ceros, desarrollado con función `zeroinfl()` de la librería `pscl`, se abordó la presencia de exceso de ceros en datos de recuentos. La función de enlace para el conteo fue "Log", indicando que los coeficientes del modelo operan en la escala logarítmica.

Se utilizó la distribución Binomial negativa para el conteo y la distribución Binomial para la inflación de ceros, caracterizando así un modelo inflado de ceros. Las métricas de ajuste, como el AIC y BIC, presentaron valores de 17552.84 y 17799.73 respectivamente. Estas métricas evalúan la bondad de ajuste del modelo, siendo preferibles valores más bajos.

El parámetro de ajuste (Theta) obtuvo un valor de 0.5512, indicando la magnitud de la inflación de ceros en el modelo. La log-verosimilitud del modelo fue -8733, y los grados de libertad asociados con la log-verosimilitud fueron 43. Estos indicadores fueron cruciales para evaluar la capacidad del modelo para describir los datos observados, considerando la presencia de ceros excesivos en la variable de interés. El proceso de optimización BFGS implicó 49 iteraciones, indicando la eficiencia del algoritmo en la convergencia del modelo durante el ajuste.

**Tabla 4.15.**  
**Resumen del modelo de regresión binomial negativa inflada de ceros.**

Información	Valor	Comentario
Tipo de modelo	Binomial negativa inflada de ceros	Modelo para datos de recuentos con exceso de ceros
Librería	pscl	
Método	zeroinfl	zeroinfl(DMORA~,link = "logit", dist = "negbin", data=datos_train)
Función de enlace (conteo)	Log	Coeficientes del modelo en escala log()
Función de enlace (inflación cero)	Logit	Log de probabilidades, y=1 sobre y=0
Distribución (conteo)	Binomial negativa	Modelo para datos de conteo
Distribución (inflación cero)	Binomial	Modelo para datos binario
AIC	17552.84	Métrica del ajuste del modelo
BIC	17799.73	Métrica del ajuste del modelo
Theta	0.5512	Parámetro de ajuste del modelo
Log-likelihood	-8733.00	Métrica del ajuste del modelo
gL: Log-likelihood	43	gL asociados con la log-verosimilitud
Iteración, optimización BFGS	49	Numero de iteraciones

Nota: Información del modelo de regresión. Destaque las métricas AIC y BIC, donde valores más bajos indican un mejor ajuste. AIC = Criterio de Información de Akaike, BIC = Criterio de Información Bayesiano, Null deviance = Desviación nula, Residual deviance = Desviación residual. gL: null = grados de libertad nulos, gL: residual = grados de libertad residuales, gL: (Log-likelihood) = grados de libertad asociados con la log-verosimilitud. Iteración, BFGS = Número de iteración del algoritmo de optimización Broyden-Fletcher-Goldfarb-Shanno.

En el análisis del modelo binomial negativo para conteos, el intercepto fue de 1.9591, indicando que, manteniendo constantes las demás variables, la tasa de incidencia era aproximadamente 7.09 veces mayor. En relación a la edad del socio (EDSOC), el coeficiente positivo de 0.2977 sugería un aumento del 35% en la tasa de incidencia por cada unidad adicional en la edad del socio.

En cuanto a los activos (ACT), se observó una disminución del 16% en la tasa de incidencia por cada unidad adicional en la variable de activos, dado el coeficiente negativo de -0.1706. Por otro lado, la cantidad solicitada del crédito (CANT) mostró un coeficiente positivo de 0.2558, asociado a un incremento del 29% en la tasa de incidencia por cada unidad adicional en el monto del crédito. En el nivel educativo (NIVEDU [2]), el coeficiente positivo de 0.2214 sugería un aumento del 25% en la tasa de incidencia para aquellos con un nivel educativo específico.

La sociedad conyugal (SOCONY [1]) presentó un coeficiente negativo de -0.1793, indicando una disminución del 17% en la tasa de incidencia para aquellos con sociedad conyugal en comparación con los individuos sin este estado civil. Respecto a la frecuencia de pago (FRECING [2]), el coeficiente positivo de 0.2344 se asoció con un aumento del 26% en la tasa de incidencia para ciertas frecuencias de pago.

En el modelo binomial para la presencia o ausencia de ceros, la edad del socio (EDSOC), con coeficiente negativo de -0.5037 indicó una disminución del 40% en la probabilidad de tener ceros con el aumento de la edad del socio.

Para los gastos (GAST), el coeficiente positivo de 0.1382 indicó un aumento del 15% en la probabilidad de tener ceros con mayores gastos. Los activos (ACT) presentaron un coeficiente positivo de 0.3421, indicando un aumento del 41% en la probabilidad de tener ceros. En cuanto a los pasivos (PAS), el coeficiente negativo de -0.0039 sugirió una ligera disminución del 0.39% en la probabilidad de tener ceros con mayores niveles de pasivos.

Los intereses pagados (INTPAG) presentaron un coeficiente negativo de -0.2199, indicando una disminución del 22% en la probabilidad de tener ceros con mayores intereses pagados. La sociedad conyugal (SOCONY [1]) mostró un coeficiente positivo de 0.3334, señalando un aumento del 33% en la probabilidad de tener ceros en comparación con aquellos sin sociedad conyugal.

El AIC, con un valor de 17552.843, respalda que el modelo se ajusta bien sin ser demasiado complejo. Además, el log-Likelihood de -8733.422 indicó que el modelo tiene una buena capacidad para explicar la realidad de los datos.

**Tabla 4.16.**  
**Parámetros del modelo de regresión binomial negativa inflado con ceros.**

Predictores	$\beta$	IRR ( $e^\beta$ )	IC	F	p
<b>Modelo de recuento</b>					
(Intercepto)	1.9591	7.09	1.54 – 32.59	2.52	<b>0.012*</b>
EDSOC	0.2977	1.35	1.03 – 1.76	2.16	<b>0.03*</b>
ING	-0.0880	0.92	0.77 – 1.09	-0.97	0.334
GAST	0.0816	1.08	0.90 – 1.31	0.85	0.395
ACT	-0.1706	0.84	0.77 – 0.92	-3.76	<b>&lt;0.001***</b>
PAS	-0.0005	1	1.00 – 1.00	-0.55	0.583
CANT	0.2558	1.29	1.09 – 1.53	2.94	<b>0.003**</b>
INTPAG	0.0866	1.09	0.98 – 1.21	1.61	0.107
GENSOC [2]	-0.0030	1	0.86 – 1.16	-0.04	0.969
NIVEDU [2]	0.2214	1.25	1.05 – 1.48	2.54	<b>0.011*</b>
NIVEDU [3]	0.0694	1.07	0.71 – 1.61	0.34	0.737
SOCONY [1]	-0.1793	0.84	0.71 – 0.99	-2.11	<b>0.035*</b>
NUMCAR [2]	-0.0158	0.98	0.82 – 1.18	-0.17	0.866
NUMCAR [3]	0.0105	1.01	0.85 – 1.20	0.12	0.905
PLAZ [2]	0.0600	1.06	0.87 – 1.30	0.58	0.562
PLAZ [3]	0.0473	1.05	0.79 – 1.40	0.32	0.747
TIPINT [1]	-0.0763	0.93	0.76 – 1.13	-0.76	0.449
REG [2]	-0.0192	0.98	0.78 – 1.23	-0.16	0.869
FRECING [2]	0.2344	1.26	1.01 – 1.59	2.02	<b>0.044*</b>
FRECING [3]	0.0220	1.02	0.83 – 1.25	0.21	0.833
ESTCLIM [2]	0.1031	1.11	0.97 – 1.27	1.46	0.143
<b>Modelo inflado a cero</b>					
(Intercepto)	-0.3293	0.72	0.07 – 7.49	-0.28	0.783
EDSOC	-0.5037	0.6	0.39 – 0.93	-2.29	<b>0.022*</b>
ING	0.1382	1.15	0.87 – 1.51	0.99	0.321
GAST	-0.3817	0.68	0.51 – 0.91	-2.63	<b>0.008**</b>
ACT	0.3421	1.41	1.23 – 1.61	4.89	<b>&lt;0.001***</b>
PAS	-0.0039	1	0.99 – 1.00	-2.9	<b>0.004**</b>
CANT	0.0999	1.11	0.87 – 1.40	0.83	0.405
INTPAG	-0.2199	0.8	0.70 – 0.92	-3.23	<b>0.001**</b>
GENSOC [2]	-0.1502	0.86	0.68 – 1.09	-1.25	0.212
NIVEDU [2]	0.0739	1.08	0.82 – 1.41	0.54	0.588
NIVEDU [3]	-0.0595	0.94	0.48 – 1.85	-0.17	0.863
SOCONY [1]	0.3334	1.4	1.07 – 1.82	2.47	<b>0.014*</b>
NUMCAR [2]	-0.0621	0.94	0.70 – 1.25	-0.42	0.672
NUMCAR [3]	-0.0876	0.92	0.69 – 1.21	-0.62	0.538
PLAZ [2]	-0.1365	0.87	0.65 – 1.18	-0.89	0.374
PLAZ [3]	0.0962	1.1	0.73 – 1.66	0.46	0.647
TIPINT [1]	0.0831	1.09	0.80 – 1.48	0.53	0.596
REG [2]	-0.2595	0.77	0.54 – 1.11	-1.41	0.158
FRECING [2]	0.1206	1.13	0.78 – 1.64	0.63	0.526
FRECING [3]	0.1989	1.22	0.87 – 1.71	1.16	0.248
ESTCLIM [2]	-0.1248	0.88	0.71 – 1.10	-1.12	0.261
Observaciones	2302				
AIC	17552.843				
log-Likelihood	-8733.422				

Nota: Resumen de estadísticas del modelo de regresión para validación.  $\beta$  = Coeficientes estimados, IRR = Incidence Rate Ratio (*Razón de Tasas de Incidencia*), IC = Intervalo de confianza, F = Estadística F de Fisher, p = Valor p (probabilidad). Niveles de Significancia: 0 \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05 \* 0.1 ' ' 1.

#### **4.5. Comparar el desempeño de cada uno de los modelos predictivos, utilizando una función de pérdida adecuada.**

En el análisis comparativo de los modelos de regresión Poisson (POI), binomial negativa (BN), Poisson inflado de ceros (ZIP) y binomial negativa inflada de ceros (ZINB), se evaluaron diversas métricas para determinar el rendimiento de cada modelo en relación con la predicción de morosidad.

Al analizar la precisión de los modelos, se observaron los valores del Error Cuadrático Medio (MCE) cercanos, con el modelo BN ligeramente inferior, indicando un mejor ajuste general. Además, la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE) mostraron resultados similares en todos los modelos, con BN y ZIP destacando ligeramente con valores cercanos a 68.295 y 40.333, respectivamente, indicando una precisión razonable en las predicciones.

En términos de la criterio de Información de Akaike (AIC) y el criterio de Información Bayesiano (BIC), se destacó un valor significativamente inferior para el modelo ZIP, con un AIC de aproximadamente 107618.3, lo que sugiere un mejor equilibrio entre ajuste y complejidad.

El logaritmo de la función de verosimilitud (Log lik), una medida de probabilidad de los datos dados los parámetros del modelo, mostró que el modelo BN obtuvo el valor más alto, alrededor de -8846.93, indicando una mayor probabilidad de observar los datos.

En cuanto a los Grados de Libertad de verosimilitud (gl), los modelos POI y BN presentaron un menor número de grados de libertad (21 y 22, respectivamente), sugiriendo una mayor simplicidad en comparación con ZIP y ZINB.

Las métricas analizadas indicaron que el modelo ZIP se destacó con un AIC considerablemente menor, lo que sugiere un rendimiento competitivo en términos de ajuste y complejidad. Además, el modelo BN exhibió el Log lik más alto, indicando una mayor probabilidad de los datos y podría ser una opción robusta para la predicción del incumplimiento de pagos en relación con los días de morosidad.

**Tabla 4.17.**  
**Métricas de comparación de modelos de regresión POI, BN, ZIP y ZINB.**

Métricas	Modelos			
	POI	BN	ZIP	ZINB
AIC	169068.309	17737.860	107618.322	17552.843
BIC	169188.881	17864.174	107859.467	17799.729
Log lik	-84513.155	-8846.930	-53767.161	-8733.422
gl	21	22	42	43
Variables Sig.	21	8	24 (18/6)	13 (6/7)

Nota: Resumen de estadísticas de rendimiento por modelo de regresión. POI = Modelo de regresión de Poisson, BN = Modelo de regresión binomial negativa, ZIP = Modelo de regresión Poisson inflado de ceros, ZINB = Modelo de regresión Binomial negativa inflada de ceros. ECM = Error cuadrático medio, RSME = Error cuadrático medio, MAE = Error absoluto medio, AIC = Criterio de información de Akaike, BIC = Criterio de información bayesiano, Log lik = Logaritmo de la función de verosimilitud, gl = Grados de libertad de verosimilitud. Variables Sig. = variables significativas nivel de significancia del 95%.

A pesar de que los valores de AIC y BIC no son excepcionalmente bajos, podrían indicar cierta dificultad para los modelos en predecir con precisión los valores observados. Investigaciones previas han demostrado que fenómenos complejos o no lineales pueden presentar desafíos para los modelos predictivos lineales. Esto puede deberse a interacciones no lineales entre las variables predictoras y la variable objetivo, así como la presencia de efectos de umbral u otras relaciones no lineales que no son capturadas por modelos lineales estándar [55], [64].

La inclusión de variables importantes en los modelos es crucial para capturar la variabilidad y las relaciones subyacentes en los datos. Investigaciones anteriores han destacado la importancia de identificar y considerar todas las variables relevantes en el modelado predictivo. La omisión de variables importantes puede llevar a modelos sesgados y subestimar la verdadera complejidad del fenómeno estudiado [64], [65].

Estudios previos han señalado que el tamaño de la muestra puede influir en la precisión de las métricas de evaluación de los modelos. Para muestras pequeñas, los valores de AIC y BIC pueden no ser representativos y pueden estar sesgados hacia modelos más complejos. La falta de suficientes datos puede limitar la capacidad del modelo para generalizar patrones y relaciones en los datos de forma precisa [66], [67].

# Capítulo 5

## 5. CONCLUSIONES Y RECOMENDACIONES

### 5.1. Conclusiones

- El modelo de regresión binomial negativa inflada de ceros (ZINB) se destacó al utilizar solo 13 variables significativas, evidenciando su eficacia con un AIC de 17552.843.
- Los coeficientes estimados del modelo ZINB revelaron que ciertas variables tenían una influencia significativa en la morosidad de los socios. Entre ellas, la edad del socio (EDSOC) mostró una asociación positiva, donde un aumento en este nivel se tradujo en un incremento del 35% en la tasa de morosidad. Además, los activos del socio (ACT) y la presencia de una sociedad conyugal (SOCONY [1]) exhibieron asociaciones negativas, con un descenso del 16% en la tasa de morosidad cuando los activos del socio eran mayores y cuando no existía una sociedad conyugal. Asimismo, la cantidad solicitada (CANT) tuvo un impacto significativo, con un aumento del 29% en la tasa de morosidad con cada incremento en la cantidad de crédito adquirido.

### 5.2. Recomendaciones

- Realizar un análisis más detallado de la importancia de las variables predictoras para entender mejor su contribución al modelo y posiblemente reducir la complejidad del modelo si es necesario.
- El uso de técnicas estadísticas avanzadas como los Modelos GAM inflados de ceros puede ofrecer una solución efectiva para abordar problemas de morosidad en datos con excesivos ceros.
- Es conveniente considerar la inclusión de otras variables potencialmente relevantes, como el estado laboral del cliente y su historial crediticio anterior, para mejorar la capacidad predictiva del modelo.
- Se recomienda explorar posibles interacciones entre las variables predictoras para entender cómo se influyen mutuamente y cómo esto puede afectar a la variable de interés, es decir, la morosidad.



## 6. REFERENCIAS

- [1] H. Ibarra, «Movilización y organización campesina en la costa ecuatoriana (1950-1963)», *Ecuad. Debate*, vol. 80, pp. 137-148, 2010.
- [2] M. Cordova, «Proyecto de servicios de apoyo a la producción y comercialización agropecuaria de los cantones de Gualaceo, Paute, Guachapala y El Pan en la provincia del Azuay», Escuela Superior Politécnica del Litoral - Espol, 2004.
- [3] Cooperativa de Ahorro y Crédito Jardín Azuayo, *Somos Nuestra Memoria. Historia de la Cooperativa Jardín Azuayo 1996-2014*. Cuenca, Ecuador: Cooperativa de Ahorro y Crédito Jardín Azuayo, 2019.
- [4] O. Pinos, «Medición y evaluación de impacto del crédito con acompañamiento técnico colocado por la Cooperativa de Ahorro y Crédito Jardín Azuayo en Unidades Productivas Agropecuarias UPAs de los cantones nor orientales de la provincia del Azuay durante el periodo 20», Universidad de Cuenca, 2018. doi: 10.5354/0717-8883.1988.23881.
- [5] Cooperativa de Ahorro y Crédito Jardín Azuayo, «Sembrando esperanza 3», *Sembrando esperanza*, Cuenca, Ecuador, p. 80, junio de 2017.
- [6] Cooperativa de Ahorro y Crédito Jardín Azuayo, «Informe de crédito productivo COAC/JA - CECCA. Reporte de Fondos Administrados y Crédito, Marzo 2022», Azuay, Cuenca, 2022.
- [7] J. C. Trejo-García, H. Ríos-Bolívar, y M. Á. Martínez-García, «Análisis de la Administración del Riesgo Crediticio en México para Tarjetas de Crédito», *Rev. Mex. Econ. Finanz.*, vol. 11, n.º 1, pp. 103-121, 2016, doi: <https://doi.org/10.21919/remef.v11i1.79>.
- [8] J. P. Yannuzzi Rivadeneira, V. Quinde Rosales, y R. Bucaram Leverone, «El índice de orientación agrícola (IOA) como mecanismo de análisis del gasto público agropecuario ecuatoriano», *INNOVA Res. J.*, vol. 3, n.º 3, pp. 133-141, 2018, doi: 10.33890/innova.v3.n3.2018.463.
- [9] S. R. León-Bermeo y D. Y. Murillo-Párraga, «Análisis Financiero: Gestionar los riesgos en las Cooperativas de Ahorro y Crédito segmento 1», *Rev. Arbitr. Interdiscip. Koinonía*, vol. 6, n.º 12, pp. 241-262, 2021, doi: <https://doi.org/10.35381/r.k.v6i12.1289>.
- [10] A. Valle y I. Aguirre, «Panorama económica del Ecuador en tiempos de coronavirus», *Banco Central del Ecuador*, vol. 64, Quito, p. 27, octubre de 2020.
- [11] Fundación CODESPA, «Crédito agropecuario para productores de pequeña escala. El compromiso por las microfinanzas rurales del Banco ADOPEM y CODESPA», Primera edición 2017, 2017.
- [12] S. N. Burneo, «Financiamiento para el desarrollo: Inclusión financiera de las pymes en el Ecuador», en *Naciones Unidas, CEPAL, Naciones U.*, Santiago: Serie Financiamiento para el Desarrollo N° 260, 2016, p. 49.
- [13] E. Torres Navarrete *et al.*, «Financiamiento del cultivo de maíz en el cantón Mocache-Ecuador», *Rev. Amaz. Cienc. Tecnol.*, vol. 4, n.º 3, pp. 270-300, 2015.
- [14] L. A. Fuentes Méndez y J. D. Anido Rivas, «Impacto del crédito en la producción del sector agrícola en Venezuela, 1970-1999», *Visión Gerenc.*, n.º 1, pp. 26-46, 2007.

- [15] D. Rodríguez, «Análisis y evaluación del riesgo crediticio y su incidencia en la rentabilidad de la Cooperativa De Ahorro Y Crédito San Francisco Ltda. Agencia Pelileo.», Universidad Técnica de Ambato, 2014.
- [16] J. Salinas y V. Lojano, «Análisis de los factores que inciden en la gestión de riesgos financieros en las Cooperativas de Ahorro y Crédito del segmento cinco, Cuenca 2018», *Obs. Econ. Latinoam.*, vol. Febrero, n.º 2019, pp. 1-15, 2019.
- [17] J. Altuve y A. J. Hurtado Briceño, «Análisis de los factores que influyen en la morosidad del sistema bancario venezolano (2005-2015)», *Rev. Venez. Análisis Coyunt.*, vol. 19, n.º 1, pp. 59-83, 2018.
- [18] E. Spuchlřáková, K. Valařková, y P. Adamko, «The Credit Risk and its Measurement, Hedging and Monitoring», *Procedia Econ. Finance*, vol. 24, n.º July, pp. 675-681, 2015, doi: 10.1016/s2212-5671(15)00671-1.
- [19] D. A. V. De Paula, R. Artes, F. Ayres, y A. M. A. F. Minardi, «Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques», *RAUSP Manag. J.*, vol. 54, n.º 3, pp. 321-336, jul. 2019, doi: 10.1108/RAUSP-03-2018-0003.
- [20] Abdullah, D. Zhou, S. Ali Khan, K. Jebran, y A. Ali, «Agricultural Credit in Pakistan: Past Trends and Future Prospects», *J. Appl. Environ. Biol. Sci.*, vol. 5, n.º 12, pp. 178-188, 2015.
- [21] B. Balana y M. Oyeyemi, «Agricultural credit constraints in smallholder farming in developing countries : Evidence from Nigeria», *World Dev. Sustain.*, vol. 1, n.º March, p. 100012, 2022, doi: 10.1016/j.wds.2022.100012.
- [22] X. Liqiu y L. Yanqiu, «Relationship Between Agricultural Credits and Agricultural Economy Based on Error Correct Model in Heilongjiang Province», *J. Northeast Agric. Univ.*, vol. 18, n.º 1, pp. 75-78, 2011, doi: 10.1016/S1006-8104(13)60087-2.
- [23] K. Brown y P. Moles, «Credit Risk Management», en *UK: Edinburgh Business School, Heriot-Watt University.*, 2016.
- [24] Y. Ramírez y I. P. Blanco, «Caracterización de los riesgos financieros de los productores en el sector agrícolas del municipio de Sibaté Cundinamarca», Trabajo de grado, Corporación Universitaria Minuto de Dios, Bogotá, Colombia, 2020.
- [25] J. R. García, «Análisis de los principales factores de riesgo en la producción y comercialización del maíz del Cantón Baba, periodo 2013-2017», Universidad de Guayaquil, Facultad de Ciencias Económicas, Guayaquil, Ecuador, 2019.
- [26] J. P. Hughes y C.-G. Moon, «How bad is a bad loan? Distinguishing inherent credit risk from inefficient lending (does the capital market price this difference?)», New Brunswick, NJ, Working Paper, No. 2017-09, 2017. doi: <http://hdl.handle.net/10419/200270>.
- [27] J. Yucailla, «Metodología de gestión de control, seguimiento y recuperación de la cartera de crédito con calificación "E" en la Cooperativa de Ahorro y Crédito Ambato Ltda.», Universidad Tecnológica Indoamérica, 2021.
- [28] M. Naili y Y. Lahrichi, «The determinants of banks' credit risk: Review of the literature and future research agenda», *Int. J. Finance Econ.*, vol. 27, n.º 1, pp. 334-360, ene. 2022, doi: 10.1002/ijfe.2156.
- [29] R. Chavarín Rodríguez, «Morosidad en el Pago de Créditos y Rentabilidad de la Banca Comercial en México», *Rev. Mex. Econ. Finanz.*, vol. 10, n.º 1, pp. 71-83, ene. 2015, doi: 10.21919/remef.v10i1.67.

- [30] M. M. Madrid y E. J. Zambrano, «La morosidad y su incidencia en la liquidez en la pandemia del COVID 19 en una empresa de telecomunicaciones», *Uniandes Episteme*, vol. 10, n.º 1, pp. 15-28, 2023.
- [31] W. Hurtado, «El Impacto del ciclo económico sobre la tasa de morosidad del crédito consumo condicionado al tipo de entidad financiera en el Perú», Trabajo de investigación, Universidad Peruana de Ciencias Aplicadas, Facultad de Economía, Lima, Perú, 2021.
- [32] M. Figueroa, A. San Martín, y J. Soto, «Buenas Prácticas Financieras en la Gestión del Riesgo de Crédito de Empresas del Sector Bancario», Tesis, Pontificia Universidad Católica del Perú, Escuela de Posgrado, Surco, Perú, 2018.
- [33] C. González y R. Gutiérrez, «Análisis financiero de las medidas adoptadas como impulso económico por el sector bancario de Centro América en época de pandemia», *Cent. - Rev. Científica Univ.*, vol. 12, n.º 1, pp. 38-59, 2023.
- [34] Junta de Política y Regulación Monetaria y Financiera, *Norma para la gestión de riesgos de créditos en las cooperativas de ahorro y crédito*, n.º 129. Resolución No. 129-2015-F LA, 2015, pp. 1-11. doi: Resolución No. 129-2015-F LA.
- [35] Superintendencia de Economía Popular y Solidaria - SEPS, *Reforma a las Normas para la Gestión del Riesgo de Crédito y para la Constitución de Provisiones en las Cooperativas de Ahorro y Crédito y Asociaciones Mutualistas de Ahorro y Crédito para la Vivienda*. Ecuador: Oficio Nro. SEPS-SGD-IGT-2020-36013-OFC, 2020, p. 5. doi: Oficio Nro. SEPS-SGD-IGT-2020-36013-OFC.
- [36] J. Möllmann, M. Buchholz, W. Kölle, y O. Musshoff, «Do remotely-sensed vegetation health indices explain credit risk in agricultural microfinance?», *World Dev.*, vol. 127, 2020, doi: 10.1016/j.worlddev.2019.104771.
- [37] M. Cussi, «Influencia de factores macroeconómicos sobre el riesgo crediticio de la banca microfinanciera (2000 - 2017)», Tesis de grado, Universidad Mayor de San Andrés, Facultad de Ciencias Económicas y Financieras, La Paz, Bolivia, 2019.
- [38] L. A. Contreras Gómez, «Morosidad de la cartera de crédito al consumo y su incidencia en la rentabilidad y liquidez del Banco Mercantil, Banco Universal», *Gest. Desarro. Libre*, vol. 5, n.º 9, pp. 121-149, 2020, doi: 10.18041/2539-3669/gestionlibre.9.2020.8109.
- [39] E. Ali y N. E. Awade, «Credit constraints and soybean farmers' welfare in subsistence agriculture in Togo», *Heliyon*, vol. 5, n.º 4, pp. 1-24, 2019, doi: 10.1016/j.heliyon.2019.e01550.
- [40] A. H. Abdallah, «Agricultural credit and technical efficiency in Ghana: is there a nexus?», *Agric. Finance Rev.*, vol. 76, n.º 2, pp. 309-324, 2016, doi: 10.1108/AFR-01-2016-0002.
- [41] X. Hao, Q. Sun, y F. Xie, «The COVID-19 pandemic, consumption and sovereign credit risk: Cross-country evidence», *Econ. Model.*, vol. 109, n.º December 2021, p. 13, 2022, doi: 10.1016/j.econmod.2022.105794.
- [42] S. Shahab, M. M. Ahmad, S. Panezai, y I. A. Rana, «An empirical assessment of farmers' risk attitudes in flood-prone areas of Pakistan», *Int. J. Disaster Risk Reduct.*, vol. 18, pp. 107-114, 2016, doi: 10.1016/j.ijdrr.2016.06.007.

- [43] W. Akram, Z. Hussain, M. H. Sial, y I. Hussain, «Agricultural credit constraints and borrowing behavior of farmers in Rural Punjab», *Eur. J. Sci. Res.*, vol. 23, n.º 2, pp. 294-304, 2008.
- [44] A. Elizondo, *Medición integral del riesgo de crédito*. México DF, 2012.
- [45] J. O. Arteaga y W. E. Caicedo, «Análisis de variables asociadas al riesgo crédito, para la construcción de un modelo predictivo de clasificación estadístico (credit scoring) para la empresa contactar», Trabajo de investigación, Universidad de Nariño, Facultad de Ciencias Económicas y Administrativas, San Juan de Pasto, Colombia, 2014.
- [46] R. Villamil, «Modelo predictivo neuronal para la evaluación del riesgo crediticio», Tesis, Universidad Nacional de Colombia, Facultad de Ciencias Económicas, Bogotá, Colombia, 2013.
- [47] O. Wagdi y Y. Tarek, «The Integration of Big Data and Artificial Neural Networks for Enhancing Credit Risk Scoring in Emerging Markets: Evidence from Egypt», *Int. J. Econ. Finance*, vol. 14, n.º 2, pp. 32-43, 2022, doi: 10.5539/ijef.v14n2p32.
- [48] M. P. Peña y J. M. Orellana, «Red neuronal para clasificación de riesgo en cooperativas de ahorro y crédito», *Congr. Cienc. Tecnol. ESPE*, vol. 13, n.º 1, pp. 121-124, 2018, doi: 10.24133/cctespe.v13i1.710.
- [49] J. N. Crook, D. B. Edelman, y L. C. Thomas, «Recent developments in consumer credit risk assessment», *Eur. J. Oper. Res.*, vol. 183, n.º 3, pp. 1447-1465, 2007, doi: <https://doi.org/10.1016/j.ejor.2006.09.100>.
- [50] M. J. Mures Quintana, A. García Gallego, y M. E. Vallejo Pascual, «Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras: comparación de resultados», *Pecvnia Rev. Fac. Cienc. Económicas Empres. Univ. Ón*, vol. 1, n.º 1, p. 175, 2005, doi: 10.18002/pec.v0i1.746.
- [51] S. Rayo, J. Lara, y D. Camino, «A Credit Scoring Model for Institutions of Microfinance under the Basel II Normative», *J. Econ. Finance Adm. Sci.*, vol. 15, n.º 28, pp. 89-124, jun. 2010.
- [52] C. Garay, «Modelo para monitoreo, ajuste y predicción de indicadores de morosidad», Tesis, Universidad de los Andes, Facultad de Ingeniería, Bogotá, Colombia, 2014.
- [53] J. Zevallos, «Gestión de riesgo crediticio y el nivel de morosidad en las cajas municipales de la provincia de Chupaca - 2020», Tesis, Universidad Continental, Huancayo, Perú, 2022.
- [54] C. López y S. Quizhpe, «Estudio comparativo de la morosidad que presentan las Cooperativas de Ahorro y Crédito del segmento 1 y 4 en el periodo 2015-2019: Propuesta de Gestión Financiera», Trabajo de titulación, Universidad de las Fuerzas Armadas, Departamento de Ciencias Económicas, Administrativas y del Comercio, 2020.
- [55] M. García-Artilles, E. Gómez-Déniz, y N. Dávila, «El modelo Poisson generalizado inflado de ceros: una aplicación en el entorno educativo universitario», *Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA*, vol. 15, pp. 159-171, 2014.
- [56] R. Álvarez-Vaz y F. Massa, «Modelos de conteo alternativos para los componentes del CPO en una encuesta de salud bucal en Montevideo, Uruguay», *Rev. Fac. Cienc.*, vol. 10, n.º 2, pp. 105-125, 2021, doi: <https://doi.org/10.15446/rev.fac.cienc.v10n2.80743>.

- [57] P. Atoche, «Modelos de regresión con datos de conteo. Aplicación a competiciones deportivas», Trabajo Fin de Grado, Universidad de Sevilla, 2017.
- [58] O. Amat, R. Manini, y M. A. Renart, «Credit Concession through credit scoring : Analysis and application proposal», *Intang. Cap.*, vol. 13, n.º 1, pp. 51-70, 2017.
- [59] G. Periche-Delgado, E. Ramos-Farroñan, y A. Chamolí-Falcón, «La morosidad ante un confinamiento del Covid-19 en la Caja Rural de Ahorro y Crédito Raíz, Perú», *Investig. Valdizana*, vol. 14, n.º 4, pp. 206-212, 2020, doi: <https://doi.org/10.33554/riv.14.4.801>.
- [60] F. Paco, «los riesgos financieros de las Cooperativas de Ahorro y Crédito y su relación con la rentabilidad», Tesis de grado, Facultad de Ciencias Economicas y Financieras, Universidad Mayor de San Andrés, La Paz, 2004.
- [61] L. Cuccaro, M. Sangiácomo, y L. Tumini, «El crédito formal en la Argentina: un análisis con perspectiva de género», Documentos de Proyectos (LC/TS.2022/24), Comisión Económica para América Latina y el Caribe (CEPAL)/Banco Central de la República Argentina (BCRA), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL)/Banco Central de la República Argentina (BCRA), 2022.
- [62] J. García, «Comenzando con Weka: Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para la clase.», Madrid, España, enero de 2016.
- [63] A. Salinas-Rodríguez, B. Manrique-Espinoza, y S. G. Sosa-Rubí, «Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud», *Salud Pública México*, vol. 51, n.º 5, pp. 397-406, oct. 2009, doi: 10.1590/S0036-36342009000500007.
- [64] M. Velasco, «Un Modelo de Regresión Poisson Inflado con Ceros para Analizar datos de un Experimento de Fungicidas en Jitomate», Tesis, Universidad Veracruzana, Facultad de Estadística e Informática, Xalapa, Veracruz, 2008.
- [65] E. Martínez y P. Vaucher, «Una revisión de los modelos de conteo con excesos de ceros», Informe de Pasantía, Universidad de la República, Facultad de Ciencias Económicas y de Administración, Montevideo, Uruguay, 2018.
- [66] A. Montesinos, «Estudio del AIC y BIC en la selección de modelos de vida con datos censurados», Tesina, Centro de Investigación en Matemáticas, A.C., Guanajuato, 2011.
- [67] D. R. Martínez, J. L. Albín, J. C. Cabaleiro, T. F. Pena, F. F. Rivera, y V. Blanco, «El Criterio de Información de Akaike en la Obtención de Modelos Estadísticos de Rendimiento», presentado en XX Jornadas de Paralelismo, A Coruña, 2009, pp. 439-444.

## 7. APÉNDICES Y ANEXOS

### Anexo A. Características y detalles de la base de datos

Tabla A1.  
Números de créditos productivos por sucursal y territorio

Territorio	Sucursal JA	n, Sucursal	%, Sucursal	n , Territorio	%, Territorio
<b>Cañar</b>	Azogues	428	6.83%	582	9.29%
	Cañar	154	2.46%		
<b>Costa dos</b>	Duran	4	0.06%	761	12.15%
	Milagro	18	0.29%		
	Naranjal	197	3.15%		
	Troncal	542	8.66%		
<b>Costa tres</b>	Machala	92	1.47%	1082	17.28%
	Machala centro	3	0.05%		
	Pasaje	491	7.84%		
	Ponce Enríquez	477	7.62%		
	Santa Rosa	14	0.22%		
	Zaruma	5	0.08%		
<b>Costa uno</b>	Guayaquil	1	0.02%	1	0.02%
<b>Cuenca</b>	Cuenca	24	0.38%	27	0.43%
	Feria Libre	2	0.03%		
	Patamarca	1	0.02%		
<b>Jubones dos</b>	Nabón	6	0.10%	6	0.10%
<b>Jubones uno</b>	Oña	337	5.38%	828	13.22%
	Saraguro	491	7.84%		
<b>Oriente</b>	Gualaquiza	1	0.02%	80	1.28%
	Huamboya	27	0.43%		
	Macas	29	0.46%		
	San Juan Bosco	1	0.02%		
	Sucúa	22	0.35%		
<b>Paute</b>	Chordeleg	53	0.85%	2895	46.23%
	Gualaceo	651	10.40%		
	Palmas	561	8.96%		
	Paute	657	10.49%		
	Sevilla de Oro	253	4.04%		
	Sigsig	720	11.50%		
<b>Total</b>		<b>6262</b>	<b>100.00%</b>	<b>6262</b>	<b>100.00%</b>

Nota: Porcentajes expresados por oficina y territorio de la Cooperativa de Ahorro y Crédito Jardín Azuayo. n = Frecuencia; % = Porcentaje; JA = Cooperativa Jardín Azuayo. El conjunto de datos presentado analizaba la distribución de los préstamos solicitados en distintas sucursales de la Cooperativa de Ahorro y Crédito "Jardín Azuayo", reflejando una heterogeneidad significativa en la obtención de créditos.

**Tabla 5.1**  
**Análisis descriptivo de las variables independientes cuantitativas**

Código	Clases/Niveles	ESTcred			
		C=Cancelados		L=Liquidados	
		n	%	n	%
<b>GENsoc</b>	Femenino	1871	45.49%	839	39.59%
	Masculino	2242	54.51%	1280	60.41%
<b>NIVinst</b>	Formación intermedia	9	0.22%	4	0.19%
	No definido	2	0.05%	0	0.00%
	Primaria	2800	68.08%	1375	64.89%
	Secundaria	1050	25.53%	627	29.59%
	Sin estudios	106	2.58%	44	2.08%
	Universitaria	146	3.55%	69	3.26%
<b>ESTciv</b>	Casado/a	2123	51.62%	891	42.05%
	Divorciado/a	245	5.96%	105	4.96%
	Soltero/a	1190	28.93%	853	40.25%
	Unión libre	383	9.31%	207	9.77%
	Viudo/a	172	4.18%	63	2.97%
<b>TERR</b>	Cañar	305	7.37%	277	13.03%
	Costa dos	389	9.41%	372	17.50%
	Costa tres	670	16.20%	412	19.38%
	Costa uno	1	0.02%	0	0.00%
	Cuenca	24	0.58%	3	0.14%
	Jubones dos	6	0.15%	0	0.00%
	Jubones uno	475	11.48%	353	16.60%
	Oriente	44	1.06%	36	1.69%
<b>CALIF</b>	Paute	2222	53.72%	673	31.66%
	A1	0	0.00%	1834	86.27%
	A2	0	0.00%	95	4.47%
	A3	0	0.00%	58	2.73%
	B1	0	0.00%	10	0.47%
	B2	0	0.00%	10	0.47%
	C1	0	0.00%	10	0.47%
	C2	0	0.00%	9	0.42%
	D	0	0.00%	5	0.24%
<b>TIPdif</b>	E	0	0.00%	95	4.47%
	Novado	93	17.42%	219	84.23%
	Reestructurado	1	0.19%	3	1.15%
	Refinanciado	1	0.19%	10	3.85%
	Represtamo	439	82.21%	28	10.77%
<b>GRUPagro</b>	Agrícolas	1724	41.68%	599	42.30%
	Agropecuarios	0	0.00%	4	0.28%
	No agropecuarios	394	9.53%	1	0.07%
	Pecuarios	2018	48.79%	812	57.34%
<b>ESTcred</b>	Cancelados	4136	100.00%	0	0.00%
	Liquidados	0	0.00%	2126	100.00%

Nota: Números de créditos por clase o categoría de las variables cualitativas. n=Frecuencias, %=Porcentaje

## Anexo B. Características y validación del conjunto de datos

### R Script B1. Tratamiento de valores ausentes

```
data %>%
  select(-CALIFICACION) %>%
  filter(complete.cases(EDAD,
                        GENERO,
                        INSTRUCCION,
                        EST_CIVIL,
                        N_CARGAS,
                        NOMBRE_SOCIO))
```

### R Script B2. Análisis de valores duplicados

```
library(dplyr)
data %>%
  summarise_all(~ sum(duplicated(.))) %>%
  t()
```

### R Script B3. Análisis de valores duplicados

```
library(dplyr)
data %>%
  distinct(COD_CREDITO, .keep_all = TRUE)
```

### R Script B4. Análisis de valores únicos

```
library(dplyr)
# Variables categoricas
data %>%
  select_if(is.character) %>%
  summarise_all(n_distinct) %>%
  t()

#variables numéricas
data %>%
  select_if(is.numeric) %>%
  summarise_all(list(
    Valores_Unicos = ~n_distinct(.),
    Min = ~min(., na.rm = TRUE),
    Max = ~max(., na.rm = TRUE))) %>%
  t()
```

### R Script B5. Gráfico de dispersión entre ingresos y gastos

```
data %>%
  select(INGRESOS,GASTOS) %>%
  ggplot(.,aes(INGRESOS,GASTOS))+
  geom_point(fill="#000066", shape=21, alpha=0.5)+
  geom_smooth(method = "lm", se = FALSE, col="#FC706C")+
```



theme\_tesis

### **R Script B6. Filtro de valores mínimos con "0"**

```
data %>%
  filter(INGRESOS>0) %>%
  filter(GASTOS>0) %>%
  filter(PLAZO>0) %>%
  filter(ACTIVOS>0) %>%
  filter(INTERES_PAGADO>0) -> data
```

### **R Script B7. Obtención de lambdas óptimos por prueba Cox-Box**

```
boxcox(lm(data$ACTIVOS ~ 1)) -> boxcoxACTIVOS
  boxcoxACTIVOS$x[which.max(boxcoxACTIVOS$y)]

boxcox(lm(data$PASIVOS ~ 1)) -> boxcoxPASIVOS
  boxcoxPASIVOS$x[which.max(boxcoxPASIVOS$y)]

boxcox(lm(data$GASTOS ~ 1)) -> boxcoxGASTOS
  boxcoxGASTOS$x[which.max(boxcoxGASTOS$y)]

boxcox(lm(data$INGRESOS ~ 1)) -> boxcoxINGRESOS
  boxcoxINGRESOS$x[which.max(boxcoxINGRESOS$y)]
```

### **R Script B8. Análisis lambda óptimos de prueba Cox-Box**

```
ggplot(lambdas_optimo, aes(y = columnas, x = lambda))+
  geom_bar(stat = "identity", fill="#000066", width = 0.5)+
  labs(y = "Columnas", x = "Valor de Lambda")+
  geom_label(aes(label=lambda), size=2.75, label.r = unit(0, "lines"))+
  geom_vline(xintercept = 0, lty="dashed", col="#000066")+
  scale_x_continuous(limits = c(-2, 2),
                    breaks = c(-2, -1, -0.5, 0, 0.5, 1, 2),
                    labels = expression(frac(1, x^2),frac(1, x),frac(1,
                    sqrt(x)),log(x),sqrt(x),x,x^2))
```

### **R Script B9. Análisis porcentajes de valores perdidos según transformaciones Box-Cox**

```
data %>%
  dplyr::select(ACTIVOS, PASIVOS, GASTOS, INGRESOS) %>%
  mutate_all(~ 1/sqrt(.+1)) %>%
  gather(key="columna", value = "valor") %>%
  group_by(columna) %>%
  summarise(num_outliers = length(boxplot.stats(valor)$out)) %>%
  mutate(p_out=(num_outliers/5358)*100,
         p_org=100-p_out)

data %>%
  dplyr::select(ACTIVOS, PASIVOS, GASTOS, INGRESOS) %>%
```

```

mutate_all(~ log(.+1)) %>%
gather(key="columna", value = "valor") %>%
group_by(columna) %>%
summarise(num_outliers = length(boxplot.stats(valor)$out)) %>%
mutate(p_out=(num_outliers/5358)*100,
       p_org=100-p_out)

data %>%
dplyr::select(ACTIVOS, PASIVOS, GASTOS, INGRESOS) %>%
mutate_all(~ sqrt(.+1)) %>%
gather(key="columna", value = "valor") %>%
group_by(columna) %>%
summarise(num_outliers = length(boxplot.stats(valor)$out)) %>%
mutate(p_out=(num_outliers/5358)*100,
       p_org=100-p_out)

```

## R Script B10.

### Análisis de Boxplot de valores originales y transformaciones Box-Cox

#### #boxplot de valores originales

```

data %>%
dplyr::select(ACTIVOS, PASIVOS, GASTOS, INGRESOS) %>%
gather(key="columna", value = "valor") %>%
ggplot(., aes(x=columna,y=valor))+
geom_jitter(shape=21, col="#b3b3b3", alpha=0.50, width = 0.25)+
geom_boxplot(col = "#000066", fill="white", width = 0.5, outlier.shape = 19,
             alpha=0.5)+
labs(x="Columnas", y="Valores")+
theme(axis.text.x = element_blank(),legend.position = "bottom")

```

#### #boxplot después de transformacion boxcox

```

data04 %>%
dplyr::select(ACTIVOS, PASIVOS, GASTOS, INGRESOS) %>%
mutate(ACTIVOS_log=log(ACTIVOS+1),
       PASIVOS_sqrt=sqrt(PASIVOS+1),
       GASTOS_log=log(GASTOS+1),
       INGRESOS_log=log(INGRESOS+1)) %>%
dplyr::select(ACTIVOS_log, PASIVOS_sqrt, GASTOS_log, INGRESOS_log) %>%
gather(key="columna", value = "valor") %>%
ggplot(., aes(x=columna,y=valor))+
geom_jitter(shape=21, col="#b3b3b3", alpha=0.50, width = 0.25)+
geom_boxplot(col = "#000066", fill="white", width = 0.5,
             outlier.shape = 19, alpha=0.5)+
labs(x="Columnas", y="Valores")+
theme(axis.text.x = element_blank(),legend.position = "bottom")

```

## Anexo C. Estructura de base de datos

### R Script D1. Construcción de estructura de base de datos

```
data %>%
  transmute(DMORA=as.integer(DMORA),
            EDSOC=as.integer(EDSOC),
            ING=as.double(ING),
            GAST=as.double(GAST),
            ACT=as.double(ACT),
            PAS=as.double(PAS),
            CANT=as.double(CANT),
            INTPAG=as.double(INTPAG),
            GENSOC=factor(GENSOC,
                          levels = c("MASCULINO", "FEMENINO"),
                          labels = c(1,2)),
            NIVEDU=factor(NIVEDU,
                          levels = c("BASICO", "INTERMEDIO", "AVANZADO"),
                          labels = c(1,2,3)),
            SOCONY=factor(SOCONY,
                          levels = c("SIN CONYUGE", "CON CONYUGE"),
                          labels = c(0,1)),
            NUMCAR=factor(NUMCAR,
                          levels = c("SIN CARGAS", "HASTA 1 CARGA",
                                      "MAS DE 2 CARGAS"),
                          labels = c(1,2,3)),
            PLAZ=factor(PLAZ,
                        levels = c("MENOS DE 24 MESES", "DE 24 A 48 MESES",
                                    "MAS DE 48 MESES"),
                        labels = c(1,2,3)),
            TIPINT=factor(TIPINT, levels = c("SIN AHORRO", "CON AHORRO"),
                          labels = c(0,1)),
            REG=factor(REG, levels = c("COSTA", "SIERRA"),
                       labels = c(1,2)),
            FREcing=factor(FREcing,
                          levels = c("INGRESO MENSUAL", "INGRESO TRIMESTRAL",
                                      "INGRESO SEMESTRAL"),
                          labels = c(1,2,3)),
            ESTCLIM=factor(ESTCLIM, levels = c("INVIERNO", "VERANO"),
                           labels = c(1,2))) -> dataModelo
```

## Anexo D. Partición de base de datos y construcción de modelos de regresión

### R Script E1. Partición de base de datos

```
#particion de base datos
set.seed(123)
indices_entrenamiento <- createDataPartition(dataModelo$DMORA,
                                              p = 0.7, list = FALSE)
datos_entrenamiento <- dataModelo[indices_entrenamiento, ]
datos_prueba <- dataModelo[-indices_entrenamiento, ]
```

### R Script E2. Construcción de modelos de regresión

```
POI <- glm(DMORA ~., family = "poisson", data = datos_entrenamiento)
summary(POI)

BN <- glm.nb(DMORA ~., data = datos_entrenamiento)
summary(BN)

mZIP <- zeroinfl(DMORA ~., link = "logit", dist = "poisson", data =
datos_entrenamiento)
summary(mZIP)

ZINB <- zeroinfl(DMORA ~., link = "logit", dist = "negbin", data =
datos_entrenamiento)
summary(ZINB)
```

### R Script E3. Evaluación de modelos en conjunto de prueba

```
# Evaluar modelos en conjuntos de prueba
pred_POI <- predict(POI, newdata = datos_prueba, type = "response")
pred_BN <- predict(BN, newdata = datos_prueba, type = "response")
pred_ZIP <- predict(ZIP, newdata = datos_prueba, type = "response")
pred_ZINB <- predict(ZINB, newdata = datos_prueba, type = "response")

mse_POI <- mean((datos_prueba$DMORA - pred_POI)^2)
rmse_POI <- sqrt(mse_POI)
aic_POI <- AIC(POI)
bic_POI <- BIC(POI)
df_POI <- df.residual(POI)
logLik_POI <- logLik(POI)
df_logLik_POI <- 21

mse_BN <- mean((datos_prueba$DMORA - pred_BN)^2)
rmse_BN <- sqrt(mse_BN)
aic_BN <- AIC(BN)
bic_BN <- BIC(BN)
df_BN <- df.residual(BN)
logLik_BN <- logLik(BN)
df_logLik_BN <- 22

mse_ZIP <- mean((datos_prueba$DMORA - pred_ZIP)^2)
rmse_ZIP <- sqrt(mse_ZIP)
aic_ZIP <- AIC(ZIP)
bic_ZIP <- BIC(ZIP)
```

```

df_ZIP <- df.residual(ZIP)
logLik_ZIP <- logLik(ZIP)
df_logLik_ZIP <- 42

mse_ZINB <- mean((datos_prueba$DMORA - pred_ZINB)^2)
rmse_ZINB <- sqrt(mse_ZINB)
aic_ZINB <- AIC(ZINB)
bic_ZINB <- BIC(ZINB)
df_ZINB <- df.residual(ZINB)
logLik_ZINB <- logLik(ZINB)
df_logLik_ZINB <- 43

cor_pearson_POI <- cor(datos_prueba$DMORA, pred_POI, method = "pearson")
cor_spearman_POI <- cor(datos_prueba$DMORA, pred_POI, method = "spearman")
cor_pearson_BN <- cor(datos_prueba$DMORA, pred_BN, method = "pearson")
cor_spearman_BN <- cor(datos_prueba$DMORA, pred_BN, method = "spearman")
cor_pearson_ZIP <- cor(datos_prueba$DMORA, pred_ZIP, method = "pearson")
cor_spearman_ZIP <- cor(datos_prueba$DMORA, pred_ZIP, method = "spearman")
cor_pearson_ZINB <- cor(datos_prueba$DMORA, pred_ZINB, method = "pearson")
cor_spearman_ZINB <- cor(datos_prueba$DMORA, pred_ZINB, method = "spearman")

library(Metrics)

mae_POI <- mae(datos_entrenamiento$DMORA, predict(POI, type = "response"))
mae_BN <- mae(datos_entrenamiento$DMORA, predict(BN, type = "response"))
mae_ZIP <- mae(datos_entrenamiento$DMORA, predict(ZIP, type = "response"))
mae_ZINB <- mae(datos_entrenamiento$DMORA, predict(ZINB, type = "response"))

#construccion de tabla
resultados <- data.frame(
  Modelo=c("POI", "BN", "ZIP", "ZINB"),
  ECM=c(mse_POI,mse_BN,mse_ZIP,mse_ZINB),
  RMSE=c(rmse_POI,rmse_BN,rmse_ZIP,rmse_ZINB),
  MAE=c(mae_POI,mae_BN,mae_ZIP,mae_ZINB),
  AIC=c(aic_POI,aic_BN,aic_ZIP,aic_ZINB),
  BIC=c(bic_POI,bic_BN,bic_ZIP,bic_ZINB),
  CorPearson=c(cor_pearson_POI,cor_pearson_BN,cor_pearson_ZIP,cor_pearson_ZINB),
  CorSpearman=c(cor_spearman_POI,cor_spearman_BN,cor_spearman_ZIP,cor_spearman_ZINB),
  df_residual=c(df_POI,df_BN,df_ZIP,df_ZINB),
  logLik=c(logLik_POI,logLik_BN,logLik_ZIP,logLik_ZINB),
  df_logLik=c(df_logLik_POI,df_logLik_BN,df_logLik_ZIP,df_logLik_ZINB))

print(resultados)

```