



Pontificia Universidad
JAVERIANA
Cali

**ANALISIS PREDICTIVO DEL IMPACTO DEL CAMBIO CLIMATICO Y EL
CONFLICTO ARMADO EN EL COMPORTAMIENTO DE LA CARTERA DE
CREDITOS AGROPECUARIOS PARA LA REGIONAL OCCIDENTE**

Johan Sebastián Valencia Cárdenas
CC. 1144196510
ID 9025144

Flavio Alejandro Castillo Perea
CC. 1234195673
ID 9025343

*Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos*

Director(a)
Isabel Cristina García Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, (MAYO) DE (2026)

FICHA RESUMEN

TÍTULO DEL PROYECTO: ANALISIS PREDICTIVO DEL IMPACTO DEL CAMBIO CLIMATICO Y EL CONFLICTO ARMADO EN EL COMPORTAMIENTO DE LA CARTERA DE CREDITOS AGROPECUARIOS PARA LA REGIONAL OCCIDENTE.

1. ÁREA DE TRABAJO: Financiero
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE(S): Johan Sebastián Valencia Cárdenas y Flavio Alejandro Castillo Perea
4. CORREO ELECTRÓNICO: josevaca@javerianacali.edu.co - flavio29@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Carrera 28B3 # 72 T 24 suroriente, Santiago de Cali Tel. (+57) 302 211 79 67, Carrera 40A # 55 – 68 suroriente, Santiago de Cali Tel. (+57) 321 633 66 09
6. DIRECTOR: Isabela Cristina García Arboleda
7. VINCULACIÓN DEL DIRECTOR: Externa
8. CORREO ELECTRÓNICO DEL DIRECTOR: icga03@gmail.com
9. CO-DIRECTOR (Si aplica): NA
10. GRUPO O EMPRESA QUE LO AVALA: Entidad Bancaria
11. OTROS GRUPOS O EMPRESAS: NA
12. PALABRAS CLAVE: Agricultura, *Machine Learning*, conflicto, cambio climático, crédito.
13. FECHA DE INICIO: mayo 2025
14. DURACIÓN ESTIMADA: 12 MESES
15. RESUMEN:

El presente proyecto desarrolló un modelo de análisis predictivo para evaluar el impacto del cambio climático y el conflicto armado en el comportamiento de la cartera agropecuaria de la banca pública colombiana, con énfasis en la región occidental del país (Valle del Cauca, Cauca y Nariño).

Esta zona se caracteriza por una alta dependencia de las actividades agropecuarias, las cuales se han visto severamente afectadas tanto por fenómenos climáticos extremos como sequías, lluvias intensas y alteraciones en los ciclos productivos como por la persistencia del conflicto armado y las economías ilegales, que generan desplazamiento forzado, inseguridad y limitaciones en el acceso a mercados.

A partir del análisis de 55.454 obligaciones de crédito, se integraron variables crediticias internas con factores externos climáticos (ola invernal, sequía, factores naturales) y de conflicto armado (hechos victimizantes, desplazamiento forzado), aplicando la metodología CRISP-DM.

Se estimaron ocho modelos predictivos (Regresión Logística Binomial, Árbol de Decisión y XGBoost, con y sin ponderación de clases, más un Logit Ordinal para severidad de mora). El XGBoost ponderado se posicionó como el modelo de mayor capacidad discriminadora (AUC-ROC = 0.7921, Kappa = 0.3253), superando en aproximadamente 14 puntos porcentuales al mejor modelo lineal.

Los principales hallazgos evidenciaron que la Ola Invernal eleva el riesgo de mora un 61% respecto a factores económicos, el desplazamiento forzado lo incrementa un 42%, y las obligaciones en Valle del Cauca presentan una probabilidad de mora un 34% superior a las del Cauca. El modelo se traduce en una arquitectura de alertas tempranas y segmentación territorial del riesgo, materializada en un dashboard interactivo en Power BI y un reporte analítico, como herramientas de apoyo a la gestión proactiva de la cartera agropecuaria en la Regional Occidente.

TABLA DE CONTENIDO

Pág.

Introducción	10
1. Contextualización del proyecto.....	11
1.1. Definición del problema.....	11
1.1.1. Planteamiento del problema	21
1.1.2. Formulación del problema	21
1.1.3. Sistematización.....	21
1.2. Objetivos	22
1.2.1. Objetivo General	22
1.2.2. Objetivos Específicos.....	22
1.3. Marco de Referencia	23
1.3.1. Marco Teórico	23
1.3.2. Antecedentes.....	25
2. Cambio climático y Conflicto armado	28
2.1. Cambio climático	28
2.2. Conflicto armado.....	33
3. Metodología y desarrollo	36
3.1. Fases	37
3.2. Entendimiento del negocio.....	37
3.3. Entendimiento de los datos	38
3.3.1. Análisis exploratorio de los datos	38
3.3.1.1 Descarga y descripción del <i>dataset</i>	39
3.3.1.2 Estadísticas descriptiva – análisis univariado y bivariado	40
3.3.1.3 PCA mixto.....	44
3.3.1.4 MCA.....	44
4. Modelado.....	45
4.1 Definición de la variable objetivo	46
4.1.1 Tratamiento del desbalance de clases.....	47
4.1.2 Variables predictoras seleccionadas	47

4.1.3 Partición <i>train/test</i>	49
4.1.4 Modelado — Regresión Logística Binomial.....	49
4.2 Modelado — Árbol de Decisión (<i>rpart</i>).....	50
4.3 Modelado — <i>XGBoost</i> (Extreme Gradient <i>Boosting</i>).....	51
4.4 Modelado — Regresión Logística Ordinal (severidad de mora).....	51
4.5 Comparación y selección del modelo.....	53
5. Evaluación y resultados del modelo predictivo.....	56
5.1 Interpretación de coeficientes.....	56
5.2 Análisis de discriminación por severidad - Modelo Ordinal.....	58
5.3 Implicaciones para la gestión de riesgo crediticio agropecuario.....	59
5.4 Limitaciones del modelo.....	60
6. Riesgo Crediticio Agropecuario (PowerBI).....	61
6.1 Página Datos: Caracterización de la Cartera.....	61
6.2 Página BINOMIAL: Resultados del Modelo Binario.....	62
6.2 Resultados del Modelo de Severidad de Mora:.....	63
7. Conclusiones Y Trabajos Futuros.....	64
7.1 Conclusiones.....	64
7.2 Trabajos Futuros.....	66
8. REFERENCIAS BIBLIOGRÁFICAS.....	68
Anexos.....	71
1. Diccionario de los datos.....	71
2. Código Python — Análisis Exploratorio de Datos y Preprocesamiento.....	75
3. Código R — Análisis de Correspondencias Múltiples (MCA).....	76
4. Código R — Modelos Logit Binomial, Árbol de Decisión y <i>XGBoost</i>	76
5. Código R — Modelo Logístico Ordinal (Severidad de Mora).....	77
6. Tabla Comparativa Interactiva de Modelos.....	77
7. <i>Dashboard</i> Interactivo de Riesgo Crediticio Agropecuario - PowerBI.....	77
8. <i>Reporte Analítico de Riesgo Crediticio Agropecuario — Regional Occidente</i>	78

LISTA DE FIGURAS

Figura 1: Mapa ubicación de las 95 oficinas en el suroccidente colombiano.....	12
Figura 2: Ilustración de clientes por departamento.....	13
Figura 3: Top 10 municipios con más solicitudes de crédito.....	14
Figura 4: Mapa de calor de municipios con mayor cantidad de créditos solicitados.....	15
Figura 5: Porcentaje total de operaciones por Banca.....	16
Figura 6: Clasificación total del tipo de productor Banca.	17
Figura 7: Destino de inversión de las operaciones.....	18
Figura 8: Clasificación solicitud de créditos por genero.	19
Figura 9: Porcentaje de calificaciones sobre los créditos.....	20
Figura 10: Medición de afectaciones climáticas en la regional occidente.....	31
Figura 11: Mapa de calor de las afectaciones por inversión.	32
Figura 12: Medición hecho victimizante.	35
Figura 13: Mapa de calor de clientes victimas.....	36
Figura 14: Metodología CRISP – DM para minería de datos [6]	37
Figura 15: Información sobre el <i>dataset</i>	39
Figura 16: Variables con valores nulos	40
Figura 17: Variables numéricas – descripción variable monto desembolsado.....	41
Figura 18: Histograma y Boxplot de la variable Monto de Desembolso.....	42
Figura 19: Variable de la base de Datos.....	43
Figura 20: Distribución de la variable objetivo: mora binaria (izquierda) y severidad ordinal (derecha).....	47
Figura 21: Distribución de la severidad de mora por departamento (izquierda) y tipo de factor externo climático (derecha).....	53
Figura 22: Comparación de métricas por modelo: AUC-ROC, <i>Balanced Accuracy</i> , <i>Sensitivity</i> (izquierda) y <i>Accuracy</i> (derecha).	54
Figura 23: Curvas ROC comparadas para los seis modelos binomiales (Logit, Árbol de Decisión y XGBoost, con y sin ponderación).	55
Figura 24: Matrices de confusión de los seis modelos binomiales: Logit, Árbol de Decisión y XGBoost (con y sin ponderación).	55
Figura 25: Odds Ratios del Modelo Logit Binomial - variables con mayor impacto sobre la probabilidad de mora. Elaboración propia.	57
Figura 26: Dashboard interactivo — caracterización de la cartera agropecuaria Regional Occidente.....	62

Figura 27: Dashboard interactivo — Página BINOMIAL: métricas comparativas de modelos binomiales, curva ROC e importancia de variables.	63
Figura 28. Página MULTIMODAL del dashboard: métricas de severidad de mora por rango temporal y comparación de modelos ordinales.	64

LISTA DE TABLAS

Tabla 1: Modelo Calificación Crediticia para comercial.....	19
Tabla 2: Modelos clásicos	23
Tabla 3: Modelos de <i>Machine Learning</i>	24
Tabla 4: Modelos de árboles de decisión.....	25
Tabla 5: Antecedentes conflicto armado	26
Tabla 6: Antecedentes cambio climático.....	27
Tabla 7: Conflicto armado reportado por el Banco	33
Tabla 8: Tabla de frecuencia variable tipo de factores externos.....	43
Tabla 9: Tabla de frecuencia variable hecho victimizante.....	44
Tabla 10: comparativo de métricas.....	54

LISTA DE ANEXOS

1. Diccionario de datos.....	69
2. Código Python — Análisis Exploratorio de Datos y Preprocesamiento.....	73
3. Código R — Análisis de Correspondencias Múltiples (MCA)	74
4. Código R — Modelos Logit Binomial, Árbol de Decisión y <i>XGBoost</i>	74
5. Código R — Modelo Logístico Ordinal (Severidad de Mora).....	75
6. Tabla Comparativa Interactiva de Modelos.....	¡Error! Marcador no definido.
7. <i>Dashboard</i> Interactivo de Riesgo Crediticio Agropecuario - PowerBI.....	¡Error! Marcador no definido.5
8. <i>Reporte Analítico de Riesgo Crediticio Agropecuario — Regional Occidente</i>	¡Error! Marcador no definido.6

Introducción

En los últimos años, la cartera agropecuaria en la regional occidente de la entidad bancaria, que comprende los departamentos del Valle del Cauca, Cauca y Nariño ha mostrado señales preocupantes de deterioro. Este fenómeno está estrechamente relacionado con factores externos como el cambio climático, que afecta directamente la productividad agrícola, y el conflicto armado, que altera la dinámica económica y social de los territorios.

La falta de herramientas analíticas que permitan anticipar estos riesgos ha limitado la capacidad del Banco para tomar decisiones proactivas que contribuyan a la sostenibilidad financiera y social del crédito agropecuario.

En este contexto, los modelos predictivos se utilizan con el propósito de anticipar comportamientos aún no observados, dentro del enfoque conocido como la primera cultura de la modelización estadística. Esta se enfoca en desarrollar algoritmos eficaces que permitan construir modelos capaces de realizar predicciones precisas sobre eventos futuros, apoyando la toma de decisiones estratégicas en escenarios complejos y dinámicos.

Este proyecto propuso el diseño y desarrollo de un modelo predictivo que permita anticipar el deterioro de la cartera agropecuaria a partir de variables relacionadas con el clima, la seguridad y otros factores socioeconómicos. Para ello, se hizo uso de técnicas de ciencia de datos, aprendizaje automático y análisis exploratorio, apoyados en fuentes de información internas y externas que brindaron una visión integral del comportamiento de la cartera y sus determinantes. Se espera que el modelo contribuya a mejorar la gestión del riesgo crediticio, permitiendo una toma de decisiones más informada, focalizada y oportuna.

Además, se buscó que la herramienta sea replicable en otras regiones del país, adaptándose a sus particularidades. Los resultados esperados incluyen una segmentación de riesgo más precisa, alertas tempranas sobre potencial deterioro y recomendaciones estratégicas para mitigar impactos negativos.

1. Contextualización del proyecto

1.1. Definición del problema

La principal entidad financiera rural del país cuenta con más de 796 sedes a nivel nacional, atendiendo el sector crediticio agropecuario colombiano, considerado un motor clave para el desarrollo rural y la seguridad alimentaria.

El agro colombiano ha venido enfrentando una creciente vulnerabilidad ante factores externos que han afectado directamente su productividad, la estabilidad financiera y la seguridad alimentaria [1]. Entre estos factores se encontraron manifestaciones del cambio climático, como sequías prolongadas, lluvias intensas y fenómenos extremos, además del conflicto armado, que generó desplazamientos, destrucción de infraestructura y limitaciones en el acceso a mercados, ocasionando impactos significativos en la capacidad de pago de los productores rurales.

En la actualidad, no existe una herramienta sistemática clara y predictiva que permitiese anticipar estos riesgos con base en información histórica y geoespacial. Aunque se realizaron análisis descriptivos y de tendencias, la ausencia de un modelo analítico integral que incorporara datos climáticos, sociales y crediticios limitó la capacidad de la entidad financiera para tomar decisiones proactivas y diferenciadas según el territorio.

Desde la perspectiva de la ciencia de datos, esta situación configuró un problema de modelado predictivo multivariado y geoespacial, que requirió la integración de fuentes heterogéneas de datos (estructurados y no estructurados), el uso de técnicas avanzadas como *Machine Learning* a través de modelos supervisados y LDA, además de la implementación de metodología como *CRISP-DM*, con el fin de garantizar un proceso analítico replicable, riguroso y útil para la toma de decisiones.

La entidad financiera, como intermediario especializado en el segmento agropecuario, evidenció cómo estos eventos incidieron negativamente en el comportamiento de su cartera agropecuaria [2]. Esto se reflejó en aumentos en la morosidad, refinanciaciones, desajustes de políticas de crédito y deterioro del riesgo crediticio. Sin embargo, hasta ese momento no se había contado con un modelo predictivo integral que permitiera anticipar dichos efectos de manera conjunta y proactiva, lo que limitó la capacidad del Banco para tomar decisiones estratégicas informadas y diseñar soluciones financieras adaptadas a las condiciones reales del territorio.

La zona de estudio, correspondiente a la regional occidente está integrada por los departamentos del Valle del Cauca, Cauca y Nariño, representa un entorno particularmente desafiante y relevante para este análisis. Esta región combina una alta diversidad agroecológica con una fuerte presencia de fenómenos extremos como sequías, inundaciones y variaciones bruscas de temperatura, así como una histórica incidencia de violencia armada, economías ilegales y desplazamiento forzado. Dichos factores han afectado tanto la producción como la comercialización agrícola, impactando directamente los indicadores de cartera de la entidad financiera.

La **Figura 1** presenta la ubicación geográfica de las 95 oficinas del Banco distribuidas en los departamentos del Valle del Cauca, Cauca y Nariño. Como se observa, la red de atención se concentra principalmente en los corredores urbanos y ejes viales de mayor actividad económica. Es relevante señalar que Santiago de Cali es la única ciudad que cuenta con tres sedes, consolidándose como el principal centro operativo de la Regional Occidente.

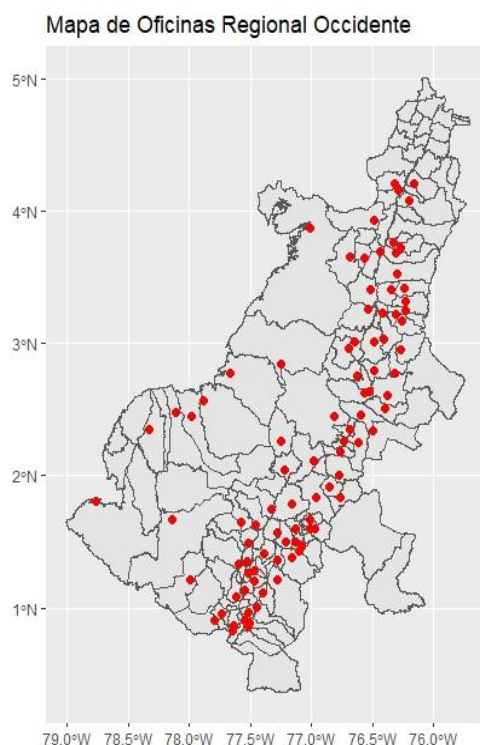


Figura 1: Mapa ubicación de las 95 oficinas en el suroccidente colombiano

En comparación con otras entidades financieras, la presencia del Banco en esta región es claramente superior. Mientras la mayoría de las entidades financieras concentra sus sucursales en las principales cabeceras municipales, el Banco opera en 95 oficinas en Valle, Cauca y Nariño, llegando a más de 70 municipios donde varios competidores no tienen ni una sola sede. Esta cobertura lo convierte en el único Banco con presencia estable en algunos de los territorios más apartados de la región, reforzando su papel como la entidad financiera con mayor alcance territorial en el suroccidente del país [13].

La **Figura 2** presenta la distribución porcentual de clientes por departamento, considerando todas las bancas. Es importante aclarar que, aunque el volumen total de registros es alto, varios clientes poseen más de una operación crediticia; por lo tanto, las cifras reflejan la concentración de operaciones por territorio y no necesariamente el número de clientes únicos. Los resultados muestran que Nariño concentra la mayor proporción (51,4%), seguido del Cauca (36%) y, en menor medida, del Valle del Cauca (12,7%).

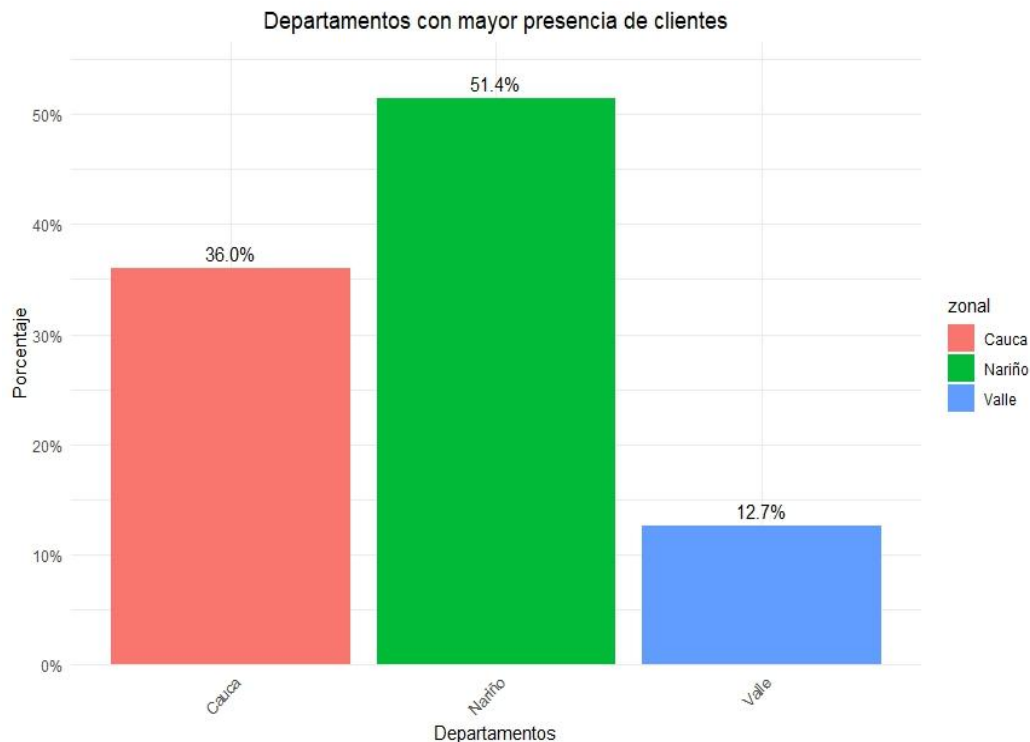


Figura 2: Ilustración de clientes por departamento

Esta distribución se explica por la dinámica económica de estos departamentos, que conforman un corredor productivo integrado, fuertemente soportado en el comercio regional y en las actividades agropecuarias, lo que genera una alta demanda de servicios financieros rurales y de productos de crédito orientados al desarrollo local.

La **Figura 3** muestra los 10 municipios con mayores solicitudes de crédito agropecuario registradas en la regional occidente. Se observa que los municipios con mayor número de solicitudes corresponden principalmente a los ubicados en el departamento de Nariño, encabezados por Ipiales, Pasto y La Unión. Estos tres municipios concentran una porción significativa de la demanda crediticia regional, lo que puede explicarse por su papel como centros de acopio y comercialización de productos agrícolas y su dinamismo agropecuario.

Municipio	Conteo de Solicitudes
IPIALES	8068
PASTO	7225
LA UNION (NARIÑO)	7186
POPAYAN SUCURSAL	6047
EL TAMBO (CAUCA)	5486
TIMBIO	5431
TUQUERRES	5369
BOLIVAR (CAUCA)	5349
CUMBAL	5115
EL BORDO	5003

Figura 3: Top 10 municipios con más solicitudes de crédito

Es importante tener en cuenta que, en el departamento del Cauca, los municipios de Popayán, El Tambo, Timbío y Bolívar muestran una participación relevante, evidenciando la importancia del crédito rural para el sostenimiento de las actividades agrícolas en territorios caracterizados por una alta diversidad productiva y una notable presencia de pequeños productores.

La **Figura 4** presenta la distribución espacial del número total de solicitudes de crédito registradas por municipio en la región de estudio. Esta representación busca evidenciar los territorios con mayor concentración de peticiones crediticias, lo que puede asociarse a la dinámica productiva, la presencia institucional o las características socioeconómicas locales.

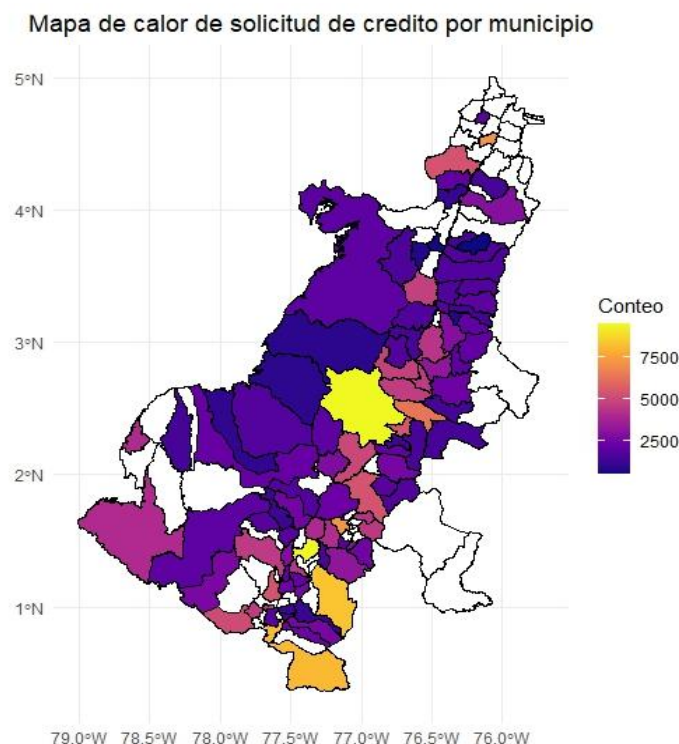


Figura 4: Mapa de calor de municipios con mayor cantidad de créditos solicitados

La distribución observada revela, la profundidad que presenta el Banco al llegar a zonas de alta complejidad en diferentes maneras (acceso, violencia, entre otros), que permite el acceso al crédito rural, además de cerrar brechas socioeconómicas y financieras de los productores.

La **Figura 5** muestra la distribución de la cartera de la Regional Occidente. Aunque la cartera se divide en seis bancas, la **Banca Agropecuaria** destaca ampliamente como el segmento de mayor representación, en coherencia con la misión institucional del Banco orientada al fomento del sector rural[14]. Este segmento agrupa más de **204.330 operaciones**, equivalentes al **81% del total**, por un valor superior a **1,8 billones de pesos**, lo que confirma su rol predominante dentro del portafolio y su fuerte incidencia en los indicadores globales en el suroccidente colombiano.

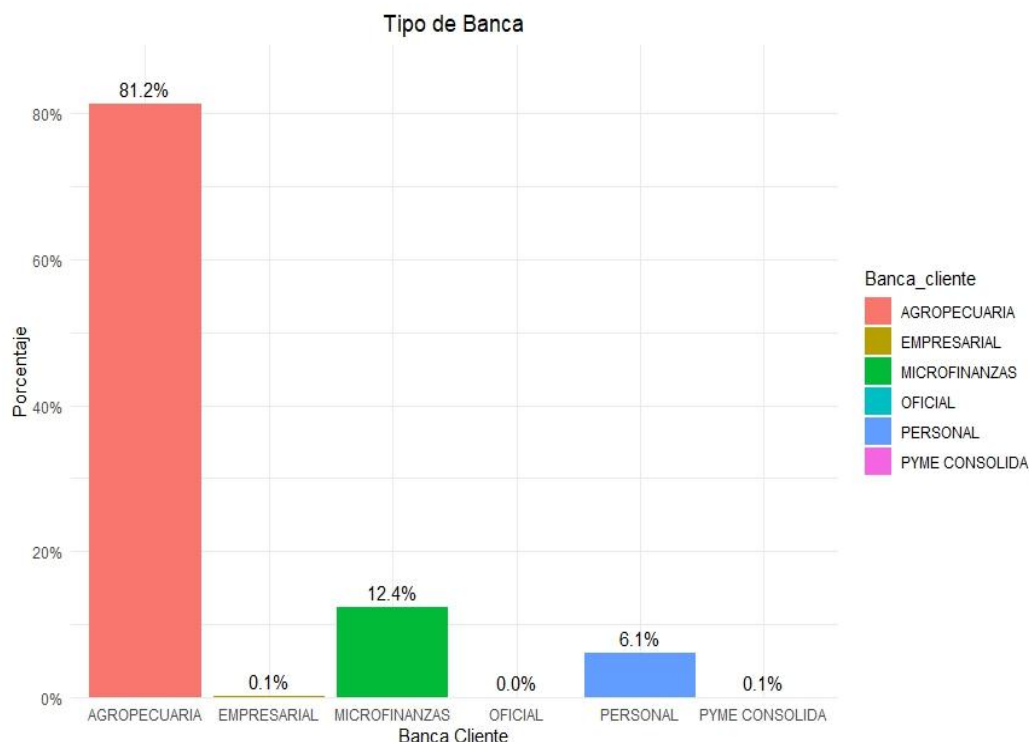


Figura 5: Porcentaje total de operaciones por Banca.

Esta elevada participación también se refleja en la distribución territorial de los recursos, con mayores niveles de inversión en los municipios de El Tambo (Cauca), Cali (Valle del Cauca), Pasto e Ipiales (Nariño), territorios donde la actividad agrícola y pecuaria es dominante. Esta concentración sectorial y geográfica constituye un elemento clave para comprender el comportamiento del riesgo crediticio y su sensibilidad frente a eventos climáticos y dinámicas del conflicto armado en la región.

La **Figura 6** discrimina la cartera por tipo de productor donde se evidencia una marcada concentración en los Pequeños Productores, quienes registran 187.367 obligaciones, equivalentes al 74% del total. Este comportamiento refleja no solo la estructura productiva de la región, sino también la orientación institucional del Banco hacia el fomento rural. Les siguen los segmentos de Microfinanzas (30.733 obligaciones) y No Agropecuario (30.260), que en conjunto aportan cerca del 24% del portafolio, consolidándose como grupos relevantes, aunque claramente secundarios frente al peso del pequeño productor.

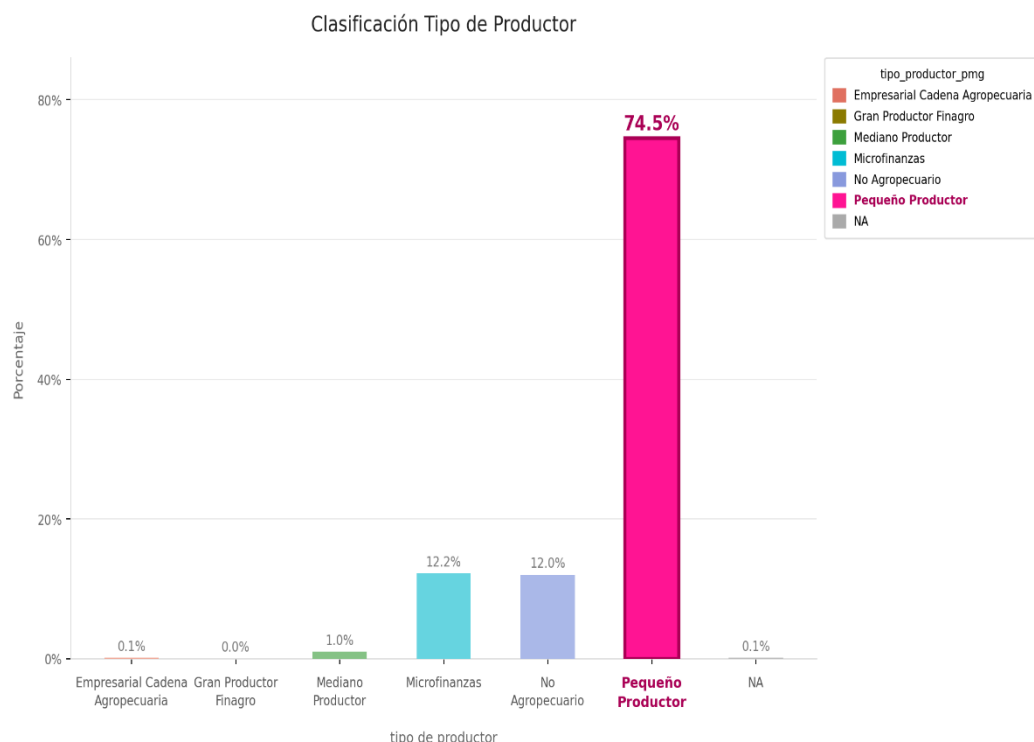


Figura 6: Clasificación total del tipo de productor Banca.

Es importante tener en cuenta que los demás segmentos como medianos Productores, Empresarial, Cadena Agropecuaria y Grandes Productores Finagro, tienen una participación marginal, cada uno con menos del 2% de la cartera. Esta distribución confirma el enfoque social y territorial de la entidad, cuyo accionar prioriza la inclusión financiera en zonas rurales y el impulso a pequeños productores, actores fundamentales para la dinámica agrícola y el desarrollo económico local.

La **Figura 7** muestra los destinos más relevantes dentro de los créditos otorgados por el Banco, evidenciando una marcada concentración en actividades agropecuarias. El destino con mayor participación es Café, que representa el 24,9% del total de obligaciones (62.647 créditos). Le siguen Ganadería con 14,6% y Microfinanzas con 12,2%, mientras que Consumo y Caña panelera aportan 5,5% y 4,4%, respectivamente. En conjunto, los principales destinos agropecuarios superan el 40% del portafolio total, lo que demuestra el peso estructural del sector agro en las regiones atendidas por el Banco y su rol estratégico en el desarrollo rural del país.

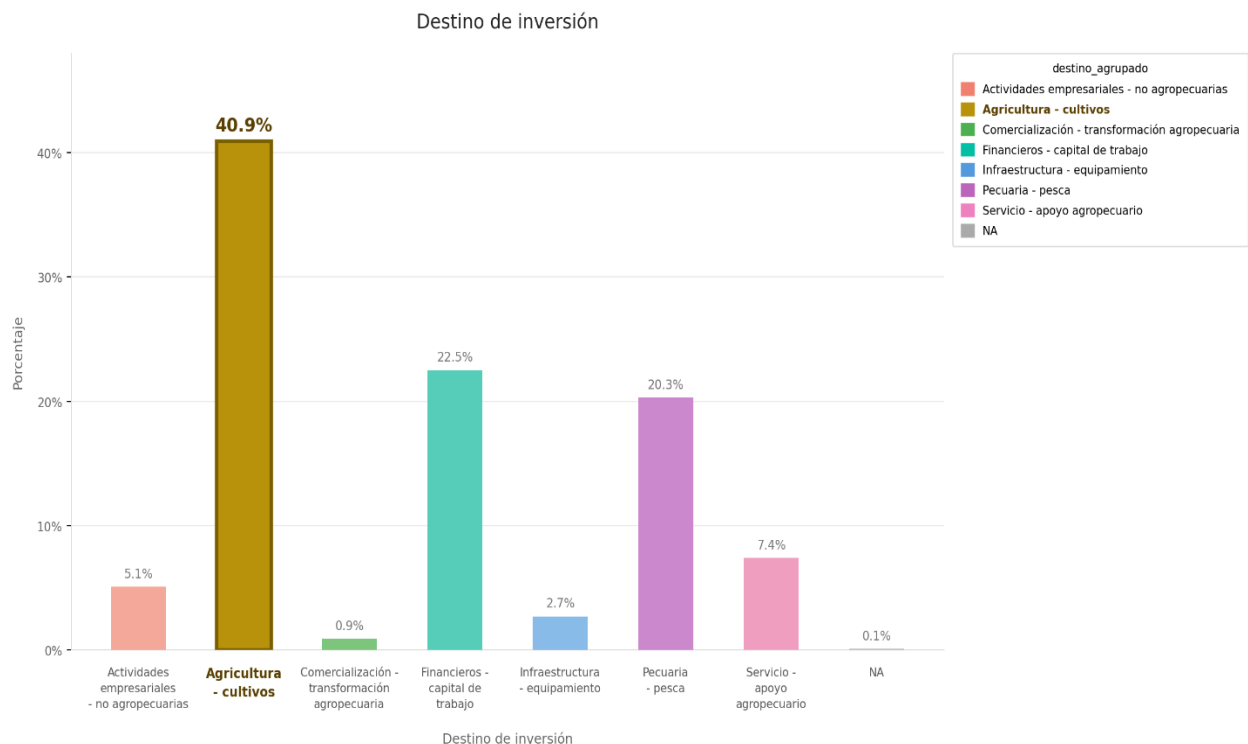


Figura 7: Destino de inversión de las operaciones.

Al comparar esta distribución con los destinos no agropecuarios, se evidencia una participación significativamente menor. Mientras los destinos agro concentran más del 40% de las obligaciones, los destinos no agro como Microfinanzas (12,2%), Consumo (5,5%) y Vivienda (3,9%) muestran un peso más reducido dentro del portafolio. Esto indica que, aunque los créditos no agro cumplen un papel importante en la diversificación y en el apoyo al comercio, la vivienda y el bienestar de los hogares, su aporte es claramente inferior frente al núcleo productivo agropecuario. De esta forma, la estructura del portafolio confirma la especialización del Banco en el sector rural y su liderazgo como principal financiador del agro colombiano.

La **Figura 8** presenta la distribución por género de los clientes con obligaciones de crédito. Si bien la participación masculina es mayoritaria con 134.079 obligaciones, equivalentes al 53% del total, destaca la relevancia creciente de las mujeres rurales en el acceso al financiamiento formal, quienes concentran 116.152 obligaciones, es decir, aproximadamente el 46% de la cartera.

Este peso femenino es especialmente significativo si se tiene en cuenta que, en departamentos como Valle, Cauca y Nariño, gran parte de la actividad productiva rural está ligada a unidades agropecuarias de pequeña escala, donde históricamente el acceso al crédito para mujeres ha sido más limitado. El comportamiento observado sugiere un avance importante en inclusión financiera, particularmente en programas dirigidos a agricultura familiar, microempresas rurales y actividades asociadas a transformación artesanal.

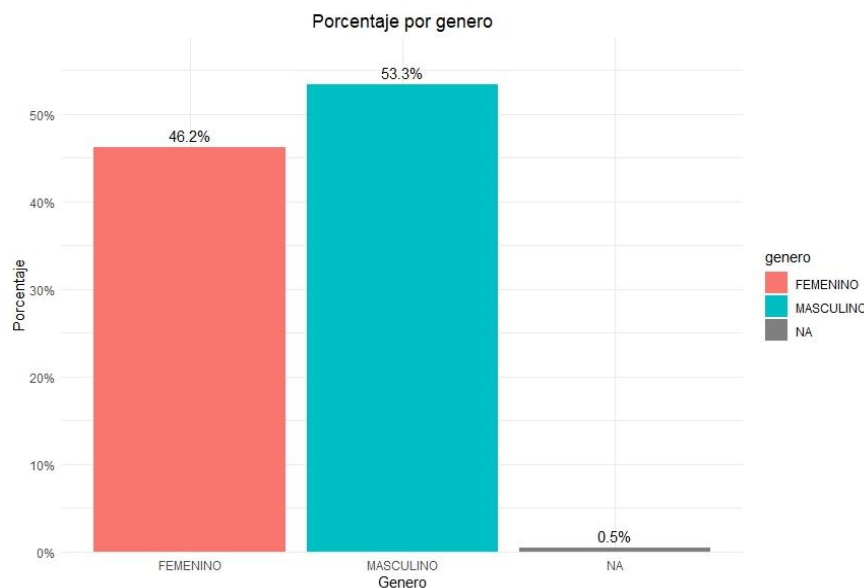


Figura 8: Clasificación solicitud de créditos por genero.

A pesar de que la cartera muestra una participación importante de mujeres en el acceso al crédito, persisten brechas estructurales que ayudan a contextualizar esta distribución. Según la nota estadística sobre la Mujer Rural del DANE [26], la tasa de aprobación de créditos para mujeres rurales es del 49,8 %, inferior a la de los hombres (56,8 %), mientras que la tasa de rechazo en crédito comercial alcanza el 88,4 % para mujeres frente al 69,2 % para hombres. Estas cifras evidencian que, aunque la participación femenina en su cartera es significativa, las mujeres rurales continúan enfrentando mayores barreras de acceso al financiamiento formal.

En la **Tabla 1** se muestra un análisis del modelo de calificación (homologado) utilizado por el Banco. Es importante señalar que los resultados de esta calificación interna no dependen únicamente de los días de mora, sino también de diversas métricas propias de la entidad. Entre ellas se destaca el control de inversión, una herramienta diseñada para realizar seguimiento a los créditos agropecuarios. Este indicador evalúa, en términos porcentuales, el grado de cumplimiento y la adecuada utilización de los recursos desembolsados durante la ejecución del proyecto productivo, constituyéndose en un insumo clave para la valoración del riesgo y la asignación final de la calificación del cliente.

DIAS DE MORA	CALIFICACIÓN MRC	HOMOLOGADA
0 - 29	AA	A
30 – 59	A	B
60- 89	BB	B
90-119	B	C
120 – 149	CC	C
>149	I	D o E

Tabla 1: Modelo Calificación Crediticia para comercial

Es importante tener en cuenta, que según el Modelo de Referencia Comercial (MRC) definido por la Superintendencia Financiera [20]. Aunque este organismo establece los lineamientos generales, el Banco cuenta con un sistema interno de calificación que, para efectos de este análisis, se presenta en su versión homologada. Esta calificación puede asignarse por obligación específica o de manera consolidada a nivel de cliente.

La **Figura 9** presenta una descripción de la calificación interna del banco donde A es la mejor y E la peor.

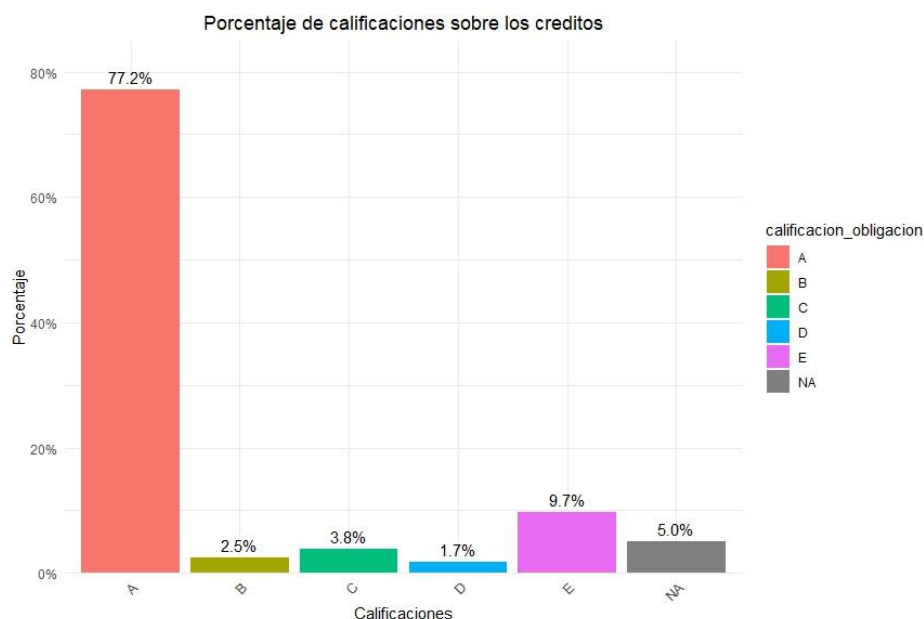


Figura 9: Porcentaje de calificaciones sobre los créditos.

La clasificación del MRC es un elemento fundamental en los procesos de evaluación crediticia, ya que influye directamente en la aprobación y condiciones de los créditos. Este aspecto cobra especial relevancia en municipios de difícil acceso, donde la evaluación del riesgo crediticio es más sensible y determinante para la inclusión financiera.

En este contexto, el uso de la analítica predictiva se convirtió en una herramienta estratégica para la gestión de riesgos crediticios. Se aplicaron algoritmos de clasificación, regresión. Con estas herramientas se logró anticipar el deterioro de la cartera en función de patrones climáticos y sociopolíticos. Asimismo, los resultados se visualizaron mediante tableros interactivos, mapas de calor y sistemas de alertas tempranas, lo que permitió a los tomadores de decisiones focalizar intervenciones, diseñar productos financieros resilientes y priorizar zonas críticas.

La falta de una solución basada en los datos representó no solo una oportunidad de innovación técnica, sino también una brecha estratégica que pudo cerrarse mediante un modelo replicable y

escalable. Este proyecto buscó precisamente llenar ese vacío, sentando las bases para un sistema de predicción temprana del riesgo agropecuario, con potencial de adaptación a otras regiones del país y a otros tipos de crédito.

1.1.1. Planteamiento del problema

La principal entidad financiera rural del país ha sido un actor clave para el desarrollo agropecuario y la seguridad alimentaria, con más de 796 sedes en todo el territorio. A pesar de ello, su cartera agropecuaria se ha visto afectada por factores externos que deterioraron la productividad, la estabilidad financiera de los productores y su capacidad de pago. Entre los más relevantes se encuentran el cambio climático con sequías, lluvias intensas e inundaciones y el conflicto armado, que provocó desplazamientos, destrucción de infraestructura y restricciones de acceso a mercados.

Las dinámicas externas mencionadas anteriormente, incrementan la morosidad, la refinanciación de créditos y el deterioro del riesgo crediticio. No obstante, la entidad no contó con un modelo predictivo integral capaz de anticipar estos efectos de manera conjunta. Los análisis realizados fueron principalmente descriptivos, lo que limitó la capacidad de tomar decisiones proactivas y diferenciadas según las condiciones territoriales.

Desde la ciencia de datos, este escenario configuró un problema de modelado predictivo multivariado y geoespacial, que requiere integrar fuentes climáticas, sociales y crediticias mediante técnicas avanzadas de *Machine Learning* y metodologías de proyectos como *CRISP-DM*. En la regional occidente (Valle del Cauca, Cauca y Nariño), la situación fue más crítica por su diversidad agroecológica y la alta exposición tanto a fenómenos climáticos extremos como a la violencia armada, lo que afectó directamente los indicadores de cartera.

En consecuencia, el problema central radicó en la ausencia de un modelo analítico integral que permitiera anticipar los riesgos crediticios asociados al cambio climático y al conflicto armado en la regional occidente. Este vacío limitó la toma de decisiones estratégicas y la capacidad de la entidad financiera para diseñar soluciones adaptadas a las condiciones reales del territorio y de los clientes.

1.1.2. Formulación del problema

¿Cómo influyen el conflicto armado y los cambios climáticos extremos en el deterioro de la cartera de créditos agropecuarios en la Regional occidente, y cómo pueden estos factores mejorar su predicción?

1.1.3. Sistematización

Para responder a la pregunta general, se plantean las siguientes sub-preguntas de investigación:

1. ¿Cuáles fueron las variables que representaron adecuadamente el conflicto armado (CERAC) y los cambios climáticos (IDEAM) y su relación con el incremento del indicador de cartera en la regional occidente?

2. ¿Qué modelos predictivos permitieron pronosticar con mayor precisión el deterioro de la cartera crediticia que involucren el número de eventos violentos y el índice de riesgo climático?
3. ¿Qué métricas de validación se utilizaron para evaluar el desempeño y la efectividad de los modelos de predicción?
4. ¿De qué manera la integración de visualizaciones y reportes analíticos, mejoró la comprensión del riesgo crediticio para la toma de decisiones en los territorios?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo predictivo que identifique y cuantifique la asociación estadística entre los factores climáticos extremos y el conflicto armado con el comportamiento de mora en la cartera de créditos agropecuarios de la Regional Occidente (Valle del Cauca, Cauca y Nariño), con el fin de segmentar el riesgo crediticio territorial y proporcionar insumos analíticos que apoyen la toma de decisiones proactiva en la gestión de cartera.

1.2.2. Objetivos Específicos

1. Analizar información histórica relacionada con el cambio climático y el conflicto armado, para identificar las zonas más afectadas en donde opera el Banco en la regional occidente (Valle del Cauca, Cauca y Nariño).
2. Implementar modelos predictivos para evaluar el impacto del cambio climático y el conflicto armado en la cartera agropecuaria.
3. Evaluar los modelos predictivos mediante métricas estadísticas de validación, con el fin de seleccionar el modelo más eficiente
4. Elaborar un reporte analítico y un *DashBoard* interactivo, que permita brindar información pertinente para la gestión proactiva del riesgo crediticio, anticipándose a las crisis derivadas de fenómenos climáticos o de seguridad.

1.3. Marco de Referencia

1.3.1. Marco Teórico

A continuación, se presenta una revisión detallada de la base teórica de los modelos que sustentan este proyecto de investigación, se hace énfasis en los modelos de riesgo crediticio, estadística y aprendizaje automático, mediante la metodología estructurada CRISP – DM, que facilite su comprensión y el modelado de los datos.

1.3.1.1 Predicción mediante modelos clásicos

Los modelos clásicos, son usados comúnmente en el mundo financiero para determinar la probabilidad de impago de una determinada solicitud (*scoring* crediticio), normalmente llamados modelos econométricos y buscan mediante variables explicativas estimar y predecir las relaciones causales o los efectos marginales, basándose en supuestos estadísticos fuertes como linealidad, normalidad de los errores, no multicolinealidad, entre otros [15].

A continuación, la **Tabla 2** agrupa los principales modelos clásicos de pronósticos y su relación con la predicción de la cartera crediticia, sumándole variables externas como el conflicto armado y el cambio climático:

Tabla 2: Modelos clásicos

Modelo	Descripción del modelo
Modelo de regresión lineal (OLS)	Su objetivo es estimar el efecto de variables explicativas sobre la variable dependiente, asumiendo que existe una relación lineal. Permite cuantificar el cambio esperado en la variable dependiente ante una variación marginal en cada regresor, bajo los supuestos de linealidad, independencia de los errores, homocedasticidad y no multicolinealidad. [29]
Modelo Logit o Probit	Se emplea para estimar la probabilidad de ocurrencia de un evento binario. La especificación Probit usa una función de distribución normal, mientras que Logit utiliza la función logística. Estos modelos transforman la combinación lineal de las variables explicativas en una probabilidad entre 0 y 1. [29]

1.3.1.2 Predicción mediante modelos de *Machine Learning*

Los modelos de *Machine Learning* se presentan como una herramienta poderosa para capturar tendencias y patrones ocultos en grandes cantidades de datos. No solo buscan explicar, sino, predecir con precisión valores futuros en modelos que no requieren supuestos lineales y pueden tener interacciones complejas. A continuación, la **Tabla 3** muestra las principales técnicas de modelos de *Machine Learning* basadas en Redes Neuronales:

Tabla 3: Modelos de *Machine Learning*

Modelo	Descripción
Regresión Regularizada (LASSO, Ridge, Elastic Net)	Introducen penalizaciones al tamaño de los coeficientes para mejorar la generalización del modelo y evitar el overfitting. LASSO (Least Absolute Shrinkage and Selection Operator) utiliza penalización L1 y permite la selección automática de variables; Ridge emplea penalización L2, reduciendo la varianza de los estimadores; Elastic Net combina ambas penalizaciones. [30]
Máquinas de soporte vectorial (SVM)	Explora construir un hiperplano óptimo que separe las observaciones en distintas clases (clasificación) o predigan valores continuos (regresión). Utilizan funciones kernel para transformar los datos en espacios de mayor dimensión, permitiendo capturar relaciones no lineales. [31]

1.3.1.3 Predicción mediante árboles de decisión

Han demostrado ser en el ámbito financiero ser una herramienta muy efectiva y eficaz para detectar patrones ocultos y complejos en la toma de decisiones con respecto a la detección de impagos. Estos modelos avanzados basados en árboles tienen la capacidad de optimizar la selección de variables y reducir el overfitting. A continuación, la **Tabla 4** presenta las principales técnicas aplicadas en el ámbito financiero:

Tabla 4: Modelos de árboles de decisión

Modelo	Descripción
Gradient Boosting / XGBoost	Basado en la técnica de boosting, entrena secuencialmente árboles de decisión donde cada nuevo árbol corrige los errores del anterior. XGBoost (Extreme Gradient <i>Boosting</i>) es una versión optimizada que mejora el rendimiento computacional y la precisión mediante regularización y paralelización. [30]
Random Forest	Es un método de ensamblaje que combina múltiples árboles de decisión contruidos sobre muestras aleatorias de los datos. El resultado final se obtiene por promedio (regresión) o votación (clasificación). [30]
CatBoost	Es un algoritmo de <i>boosting</i> por gradiente diseñado para manejar de forma eficiente variables categóricas sin necesidad de codificación previa. Utiliza un esquema de estadísticos ordenado que transforma las categorías en valores numéricos evitando la fuga de información, empleando árboles simétricos para mejorar la estabilidad y la velocidad de entrenamiento. [31]

1.3.2. Antecedentes

Existe una amplia cantidad de literatura acerca de los fenómenos que se tratan de explicar en este trabajo, con la particularidad que tienden a ser artículos, tesis, artículos, entre otros, que no explican conjuntamente la relación que hay entre la cartera, el conflicto armado y el cambio climático.

1.3.2.1 Conflicto armado

Tabla 5: Antecedentes conflicto armado

Estudio	Autores y año	Contribución
Informe de estabilidad financiera 2024-II	Banco de la República Año: 2024 [37]	El informe resalta que el sector agropecuario enfrenta un deterioro en su cartera, señalando que el indicador de calidad de mora (ICM) del sector en el año 2024 es el más alto de los últimos 5 años, debido principalmente al deterioro de subsectores como la ganadería, agricultura y caza y pesca, este asociado a condiciones macroeconómicas como: choques de precios internacionales, altas tasa de interés, violencia armada e impactos climáticos que afectaron la capacidad de pago de los productores rurales.
Conflicto armado en Colombia: evidencia de su impacto sobre la producción agrícola municipal	Juliana Libreros Ruiz Año: 2017 [4]	El documento rectifica la correlación negativa entre la producción agropecuaria y la cantidad de hectáreas y el conflicto armado en Colombia entre el 2000 y el 2014, argumentando que son los pequeños productores los más vulnerables, cambiando los incentivos de estos a participar en activas ilegales, esto convirtiéndose en un limitante para el pago de los créditos agropecuarios.
Conflicto armado y sistema financiero: impacto del conflicto e iniciativas para la inclusión financiera en épocas de posconflicto	Jose Orrego Correa Sebastián Uribe Orozco Año: 2018 [3]	Analizaron la relación entre el conflicto armado en Colombia y la inclusión financiera, utilizando un índice de inclusión financiera para 797 municipios entre 2005 y 2014. identifican el impacto negativo en la profundización del sistema financiero. El estudio propone iniciativas para fomentar la inclusión financiera en el contexto

		de posconflicto, como el uso de metodologías microfinancieras de préstamo grupal y el desarrollo de aplicaciones móviles para educación financiera.
Bancarización en Peligro: Conflicto Armado e Inclusión Financiera en Colombia	Jorge Andrés Escobar Año:2017 [28]	El conflicto interno armado de Colombia, de los más antiguos y complejos del mundo, viene acompañado desigualdad de oportunidades, incrementos en la pobreza y atraso económico que, acompañado de la literatura y las evidencias empíricas, han podido determinar la relación negativa existente entre este y los niveles de bancarización e inclusión financiera que disminuyen a su vez desarrollo económico.

1.3.2.2 Cambio climático

Tabla 6: Antecedentes cambio climático

Estudio	Autores y año	Contribución
Impacto de los factores de riesgo en el estado de la cartera del Banco agrario: análisis y recomendaciones	Fedesarrollo Año: 2016 [2]	Los fenómenos climáticos adversos, como sequías prolongadas, lluvias intensas, afectan los cultivos y, por ende, los ingresos de los agricultores y la calidad de la cartera, siendo ciertos productos y regiones más golpeados que otras, sin embargo, estos riesgos de pérdidas esperadas y no esperadas pueden ser mitigados mediante instrumentos financieros como derivados y reaseguros.
<i>Climate Risk and Access to</i>	Borradores de economía BARNREP Camilo Bohorquez-	Plantea que la variabilidad de las lluvias, afectan tanto la demanda como la oferta de créditos, limitando a los agricultores al acceso a financiamiento y dificultando el pago de los prestatarios, sumado a esto, la respuesta ambigua de los Bancos frente a estos riesgos, la caída de los ingresos y la oferta menor de crédito, mediante un

<i>Credit: Evidence from Colombian Potato Growers</i>	Penuela Margarita Gáfaró Karelys Guzmán-Finol Alex Perez Año: 2025 [5]	modelo de panel con efectos fijos que revalida que el aumento de 1% en la duración de la sequía, incrementa el valor del crédito en 0,89%.
Integración de riesgos climáticos y factores ASG en el proceso de otorgamiento de créditos bancarios	Asobancaria Año: 2025 [10]	La existencia de riesgos físicos y riesgos de transición tienen un impacto directo en la capacidad de pago de los deudores y en el valor de las garantías crediticias de los establecimientos de crédito en Colombia destacando que esto crea una oportunidad para fomentar productos financieros verdes, innovación en tecnologías inteligentes, subrayando que es de suma importancia la integración de estos riesgos ambientales, para aprovechar la estrategias sostenibles y mejorar la experiencia del cliente.
Un realismo no tan mágico: Pruebas de esfuerzo climático de la banca colombiana	Superintendencia Financiera de Colombia (SFC) en conjunto con Banco Mundial Año: 2021 [29]	Se analizó la importancia de incorporar el cambio climático en las pruebas de estrés (simulaciones de escenarios adversos) al sistema financiero en Colombia, señalando que la banca enfrenta riesgos físicos (sequías, inundaciones, entre otras) y riesgos transaccionales (cambios de políticas), sosteniendo que existen algunos Bancos que son más vulnerables frente a estos hechos que otros, debido a su exposición geográfica y sectorial.

2. Cambio climático y Conflicto armado

En el desarrollo de este objetivo se estructuró en tres etapas: (i) la recopilación y consolidación de fuentes de información históricas relacionadas con cambio climático y conflicto armado; (ii) la integración, depuración y normalización de los datos; y (iii) la identificación y análisis espacial del cambio climático y el conflicto armado dentro de la jurisdicción de la regional occidente.

2.1. Cambio climático

En la primera fase se realizó un proceso de recolección y consolidación de datos históricos relacionados con afectaciones climáticas. Para el componente de cambio climático, se utilizaron

los registros internos del Banco, los cuales se alimentan tanto de los reportes de la Unidad Nacional para la Gestión del Riesgo de Desastres (UNGRD) como de la variable institucional denominada “factor externo”, un indicador diseñado para identificar eventos climáticos y situaciones extraordinarias que impactan la cartera. Esta variable clasifica las afectaciones en las siguientes categorías:

Económico

Las afectaciones económicas derivan principalmente de la alta volatilidad de precios agropecuarios, influenciados por la oferta y demanda local e internacional, así como por la presencia de mercados inelásticos en productos sin sustitutos (como café, papa o lácteos), donde pequeñas variaciones en precio generan fuertes impactos en los ingresos del productor [22]. A esto se suma la competencia con países con mayor productividad agrícola y choques externos (tasas de cambio, precios internacionales), lo que incrementa el riesgo de pérdidas y afecta directamente la capacidad de pago de los clientes rurales.

Cada vez que aparece una afectación económica la entidad financiera activa una estrategia para la recuperación de los clientes que se encuentran afectados por estas situaciones excepcionales.

Ola invernal

La ola invernal ocurre principalmente durante las fases intensas del Fenómeno de La Niña, caracterizada por lluvias extremas, inundaciones, deslizamientos y suelos saturados. Según el IDEAM y la UNGRD, estos eventos afectan cultivos sensibles a la humedad (hortalizas, fríjol, papa, frutales) y generan daños en vías rurales, retraso en cosechas y pérdida de productividad. La estacionalidad de estos eventos suele coincidir con los picos bimodales de lluvia del suroccidente colombiano [16].

Sequía

La sequía, frecuentemente asociada al Fenómeno de El Niño, limita la disponibilidad hídrica para riego, afecta pastos para ganadería y genera estrés en cultivos dependientes del régimen de lluvias. En la Regional Occidente, la sequía es crítica para ganadería, caña panelera, plátano y productos transitorios; sin embargo, en cultivos como café, ciertos periodos secos pueden favorecer la floración cuando se presentan de forma controlada. Las sequías prolongadas incrementan la pérdida de cosechas y reducen la capacidad productiva rural.

Orden público

El suroccidente es una de las regiones más afectadas por presencia de grupos armados ilegales [33]. Según la UNP y el Observatorio del DANE, Cauca y Nariño concentran algunos de los mayores registros de afectaciones por orden público del país. La situación se agrava por bloqueos recurrentes en la Vía Panamericana, protestas comunitarias y la falta de vías terciarias transitables,

lo que genera retrasos logísticos, pérdidas en transporte de productos perecederos y disminución de ingresos. Esto limita el acceso a mercados y compromete directamente el pago de créditos.

Ante este panorama de vulnerabilidad socioeconómica, el marco regulatorio colombiano según la FAO [19] ha establecido mecanismos de protección para los actores rurales afectados. En particular, esta entidad financiera implementa medidas especiales de alivio financiero dirigidas a clientes que han sido víctimas del conflicto armado interno incluyendo población desplazada, víctimas de secuestro o desaparición forzada, en estricto cumplimiento de lo dispuesto por la Ley 1448 de 2011 [34]. Estas disposiciones buscan mitigar el impacto de la mora generada por la pérdida de activos o la imposibilidad de realizar actividades productivas en zonas de alta conflictividad.

Factores naturales – volcán

Los departamentos de Nariño y Cauca presentan actividad volcánica significativa: el volcán Galeras (Nariño) y el Puracé (Cauca) son dos de los de mayor vigilancia por el Servicio Geológico Colombiano. La caída de ceniza afecta pastos, cultivos, calidad del agua y salud animal, mientras que las alertas volcánicas generan restricciones de movilidad y potenciales evacuaciones. Estos eventos impactan la producción agrícola y la estabilidad económica de las comunidades cercanas.

Sequía y heladas

Las heladas, asociadas a descensos bruscos de temperatura, dañan tejidos vegetales, afectan cultivos de altura (papa, hortalizas, frutales) y disminuyen la disponibilidad de pasturas para ganadería. En el departamento del Cauca y Nariño, a pesar de su riqueza hídrica, la falta de infraestructura para almacenamiento y potabilización incrementa la vulnerabilidad ante eventos de sequía y heladas. Las temperaturas extremas generan pérdidas productivas significativas y aumentan la exposición financiera de pequeños productores.

Es importante resaltar que esta variable empezó a implementarse de manera sistemática a partir del año 2019, cuando el Banco estableció la necesidad de identificar y registrar de forma homogénea a los clientes y obligaciones afectados por fenómenos climáticos, con el propósito de segmentar adecuadamente su cartera y diseñar estrategias de atención para los productores impactados por estos eventos.

Para este estudio, se contó con 251.518 datos de clientes que contaban con productos en dicha entidad financiera, pero se analizaron solamente **55.454** obligaciones correspondientes a clientes afectados por diferentes situaciones climáticas durante los años 2024 y 2025, dentro de la jurisdicción de la Regional Occidente. Esta información constituye la base para evaluar la relación entre las afectaciones climáticas y el deterioro de la cartera agropecuaria en los departamentos del Valle del Cauca, Cauca y Nariño.

Estos datos, con resolución espacial a nivel municipal, se recopilaron datos secundarios en formato

XLSX con sus respectivas coordenadas y se estandarizaron en sistemas de referencia geográfica, para su visualización e interpretación.

La **Figura 10** presenta la distribución porcentual de las afectaciones externas registradas por el Banco durante los años 2024 y 2025 en los departamentos del Valle del Cauca, Cauca y Nariño. Los resultados evidencian que la Ola Invernal fue el factor predominante, concentrando más del 70% de los casos reportados, lo cual refleja la alta vulnerabilidad de la actividad agropecuaria frente a eventos climáticos extremos. En segundo lugar, las afectaciones de origen económico representan cerca del 27%, evidenciando las presiones derivadas de variaciones en precios, costos de producción y condiciones del mercado.

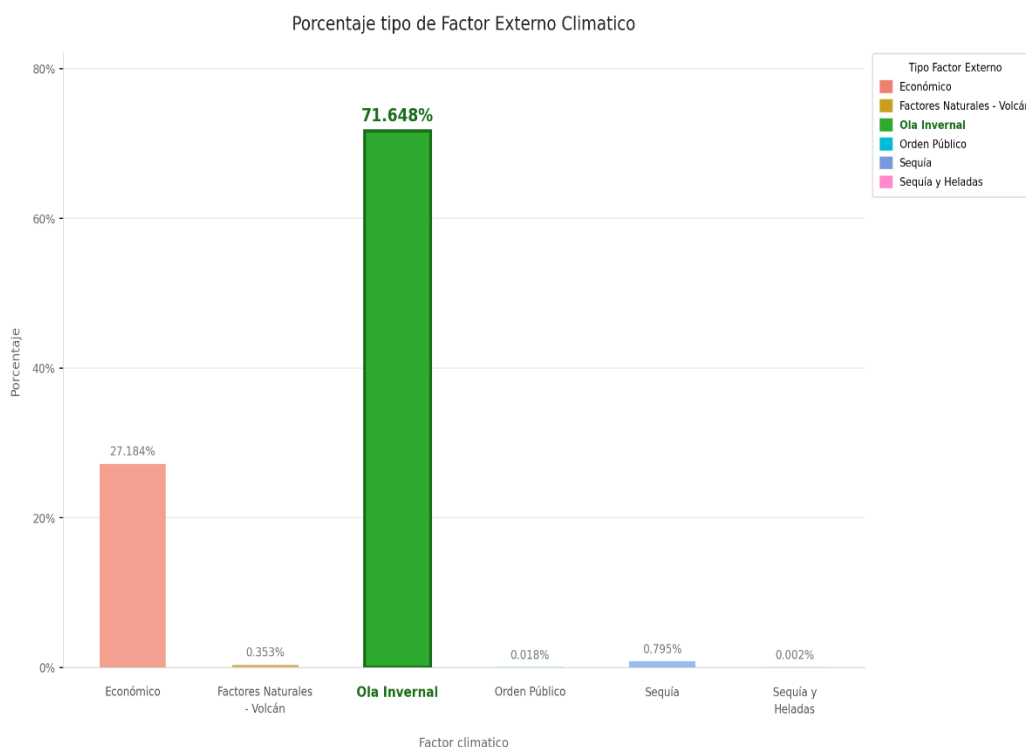


Figura 10: Medición de afectaciones climáticas en la regional occidente.

En contraste, factores como sequía, orden público y eventos naturales específicos presentan una incidencia marginal en el periodo analizado. Esta distribución permite identificar los riesgos más relevantes para la cartera agropecuaria y orienta la toma de decisiones para la gestión preventiva y la mitigación del deterioro crediticio.

La **Figura 11** analiza mediante un mapa de calor las categorías de los factores externos, permitiendo identificar la concentración geográfica de las afectaciones reportadas por el Banco en las zonas donde los clientes realizaron sus inversiones. Se observa que las categorías Económico y especialmente Ola Invernal muestran una mayor intensidad en puntos específicos del Valle del

Cauca, Cauca y Nariño, reflejando su alta frecuencia y relevancia en el deterioro de la cartera. Este patrón evidencia que ciertos municipios presentan una recurrencia significativa de eventos adversos, lo que sugiere la existencia de focos de vulnerabilidad donde las condiciones climáticas y económicas impactan de manera más pronunciada la actividad productiva. En contraste, factores como Sequía, Orden Público y Factores Naturales - Volcán muestran una dispersión menor y una frecuencia considerablemente más baja, indicando una incidencia territorial más limitada sobre todo en departamentos como Cauca y Nariño. La visualización permite, por tanto, identificar zonas críticas y orientar estrategias de mitigación y monitoreo focalizado.

Mapa de calor por categoría del Factor Externo

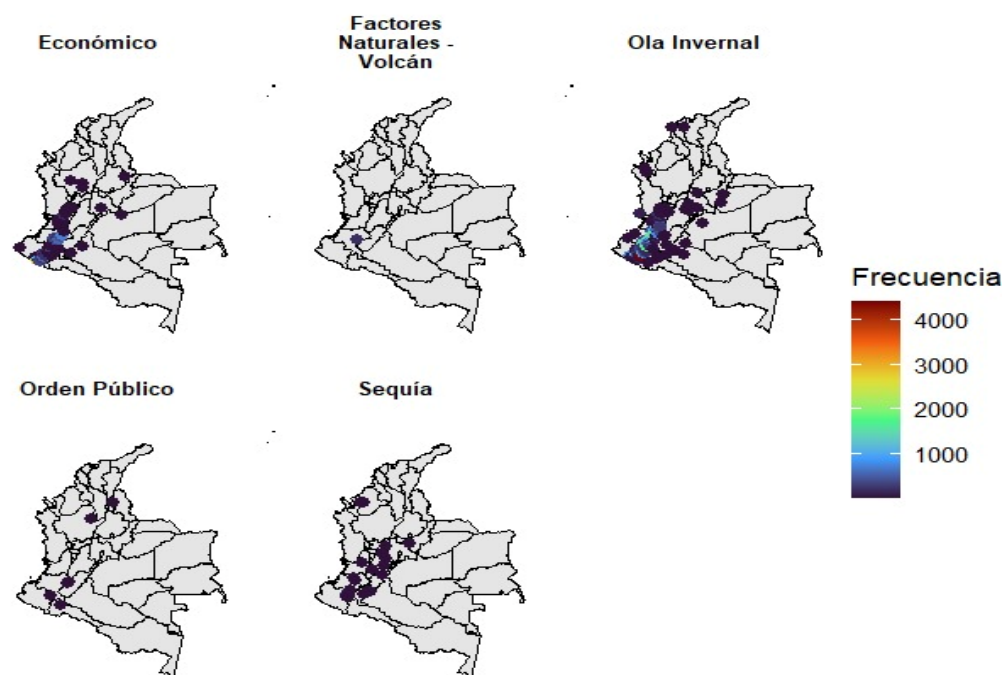


Figura 11: Mapa de calor de las afectaciones por inversión.

Es importante mencionar que los departamentos del Valle del Cauca, Cauca y Nariño se encuentran entre las zonas más vulnerables del país frente al cambio climático, según los reportes del IDEAM y la UNGRD [8], que los ubican de manera recurrente dentro de los territorios con mayor número de emergencias por inundaciones, deslizamientos, sequías y heladas en el suroccidente colombiano. Entre 2019 y 2024, la UNGRD registró en estos tres departamentos más de 2.000 eventos asociados a ola invernal y movimientos en masa, mientras que el IDEAM destaca que sus áreas rurales presentan altos niveles de exposición a la variabilidad del fenómeno El Niño y La Niña, afectando directamente cultivos como café, papa, caña panelera y hortalizas[17]. De acuerdo con el DANE [26], más del 55% de la población ocupada rural de esta región depende de actividades agropecuarias sensibles al clima, lo que incrementa la probabilidad de pérdidas productivas y

deterioro financiero cuando ocurren eventos extremos.

2.2. Conflicto armado

El conflicto armado interno en Colombia se ha expandido durante más de 6 décadas creando múltiples crisis humanitarias que han violado la integridad humana y los diferentes tratados locales e internacionales sobre los derechos humanos, donde hay zonas que han sido muy afectadas por este fenómeno que profundiza la desigualdad y la dificultad al acceso al crédito [9].

Para la variable conflicto armado, la determinación de si es víctima o no del conflicto armado esta dado por un estudio realizado por la unidad para las víctimas, donde las personas que hayan sufrido algún hecho victimizante se registran y al momento de solicitar el crédito donde el Banco hace cruces de base de datos antes, durante y después de la obtención del crédito. Esta variable contiene información detallada sobre hechos por tipo (desplazamiento forzado, homicidios, amenazas, secuestros y reclutamiento forzado), identificados por municipio e individuo, mostrada en el **Tabla 7**.

Tabla 7: Conflicto armado reportado por el Banco

Desplazamiento y despojo	
• Desplazamiento forzado • Abandono o despojo forzado de tierras • Masivo: desplazamiento forzado • Confinamiento • Otros (desplazamiento forzado, Ley 975)	Son violaciones relacionadas con la movilidad y desarraigo, uno de los hechos más masivos reconocidos en Colombia. [29]
Homicidio y desaparición	
• Homicidio • Tentativa de homicidio (Ley 975) • Desaparición forzada • Otros (desaparición forzada, Ley 975)	Entiéndase como hechos que atentan directamente contra la existencia física y la vida de las personas. [34]
Atentados, terrorismo y amenaza	
• Acto terrorista, atentados, combates, enfrentamientos y hostigamientos	La Ley los reconoce como delitos de amenaza a la vida e integridad, pero se suelen agrupar en violencia política o

• Amenaza • Masivo: actos terroristas y hostigamientos • Masivo: amenaza	bélica, con efectos colaterales en la población. [34]
Violencia física y psicológica	
• Lesiones personales físicas • Lesiones personales psicológicas • Lesiones personales y psicológicas sin incapacidad permanente • Lesiones personales y psicológicas con incapacidad permanente • Tortura	Abarca todos los daños a la integridad (personal) física y mental, con distintos niveles de gravedad, tanto colectivos como individuales. [34].
Violencia sexual y reclutamiento	
• Delitos contra la libertad e integridad sexual en desarrollo del conflicto • Vinculación de niños, niñas y adolescentes a actividades de grupos armados	Definidos como la instrumentalización del cuerpo, la infancia y adolescencia como medios de guerras. [34]
Otros hechos	
• Minas antipersonal, munición sin explotar y artefacto explosivo improvisado • Masivo: minas antipersonales • Secuestro • Otros (confinamiento, lesiones, etc.)	

La **Tabla 7** refleja la complejidad y diversidad de hechos victimizantes reconocidos en el marco del conflicto armado. Cada categoría refleja la particularidad del conflicto en Colombia y cómo todas influyen en el mismo resultado: el detrimento de las condiciones de productividad, estabilidad y seguridad. Además de estas implicaciones sociales y humanitarias, la afectación económica de la dinámica rural como: la interrupción de la producción agrícola, limitación de la movilidad y el

comercio, la reducción de la capacidad de trabajo e inversión y el acceso al crédito, afectan la cartera crediticia agropecuaria del Banco, cuya misión es orientada al desarrollo campesino y rural. Por ende, es fundamental no solo reconocer y entender estos tipos de hechos, sino también cuantificar su magnitud en las zonas afectadas donde el Banco opera, con el objetivo de comprender las alteraciones de la cartera crediticia y poder llegar a conclusiones que permitan mejorar los procesos desde el inicio del crédito teniendo en cuenta las complicaciones del conflicto que se presenta.

La **Figura 12** muestra que la mayoría de los clientes (70.5%) no reporta hechos victimizantes, lo que evidencia una alta ausencia o falta de registro de esta información. Entre quienes sí declararon afectaciones, el hecho más frecuente es el desplazamiento y despojo (24.9%), seguido por homicidio y desaparición (2.29%) y atentados o amenazas (1.10%), mientras que los demás hechos presentan porcentajes inferiores al 1%. Estas cifras indican una fuerte concentración en la categoría “No reporta” frente a una menor proporción de casos con afectación declarada.

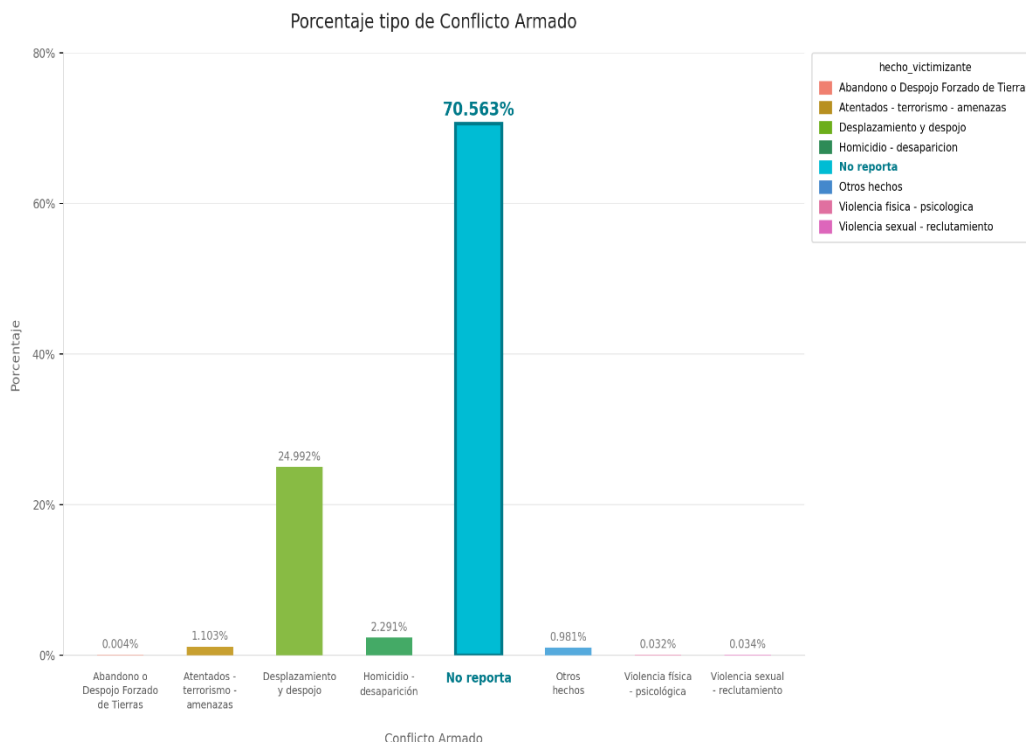


Figura 12: Medición hecho victimizante.

Los departamentos de Cauca, Nariño y Valle del Cauca continúan siendo focos críticos del conflicto armado en Colombia por la presencia de múltiples grupos armados. La **Figura 13** muestra, a través de mapas de calor, las zonas con mayor concentración de hechos victimizantes. Se observa una fuerte presencia de casos en los departamentos de Cauca y Nariño, donde la intensidad de los registros sugiere una relación significativa con las afectaciones derivadas del orden público y la persistencia del conflicto en estas regiones.

Mapa de calor por categoría de Hecho Victimizante

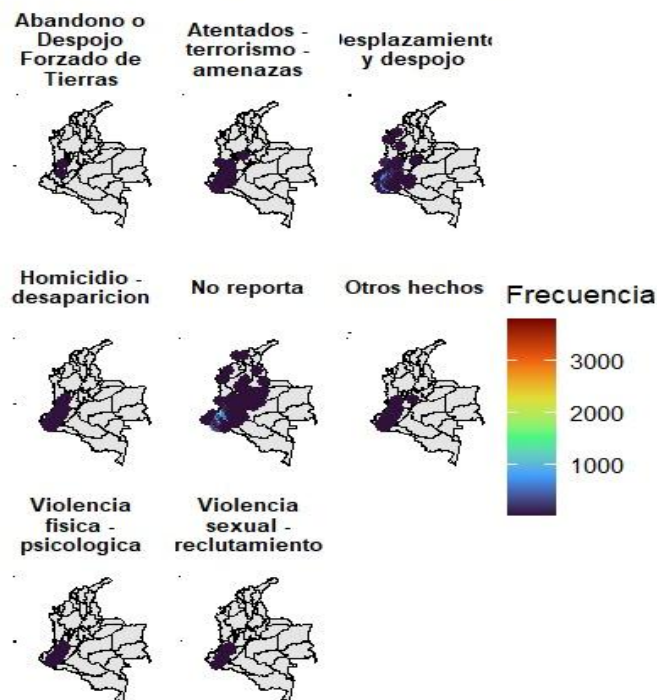


Figura 13: Mapa de calor de clientes víctimas.

En el Cauca, operan estructuras como los frentes Jaime Martínez, Dagoberto Ramos, Carlos Patiño y Adán Izquierdo, vinculadas al Estado Mayor Central (EMC), así como frentes del ELN, lo que ha generado altos niveles de violencia y reclutamiento forzado [36]. En Nariño, el EMC, junto a disidencias del ELN y otros grupos, disputan territorialidad y tráfico de economías ilícitas, con graves consecuencias para comunidades afro e indígenas incluyendo amenazas, confinamientos y reclutamiento. En el Valle del Cauca, se ha documentado una alta incidencia de violencia: el departamento registró algunas de las masacres más numerosas en 2023, según Indepaz [35], y aproximadamente el 13,8 % de su población ha sido reconocida como víctima del conflicto armado.

3. Metodología y desarrollo

Para el desarrollo de este proyecto, se adaptó la metodología CRISP – DM, método orientador de proyectos de minería de datos, donde se describen las fases de un proyecto, las tareas de cada fase necesarias para alcanzar los objetivos y la explicación de la relación de entre tareas y fases, que ofrece un resumen del ciclo de vida de un proyecto. Su aplicación permitió plantear de manera sistemática el problema, siguiendo sus seis fases principales de la **Figura 14** :

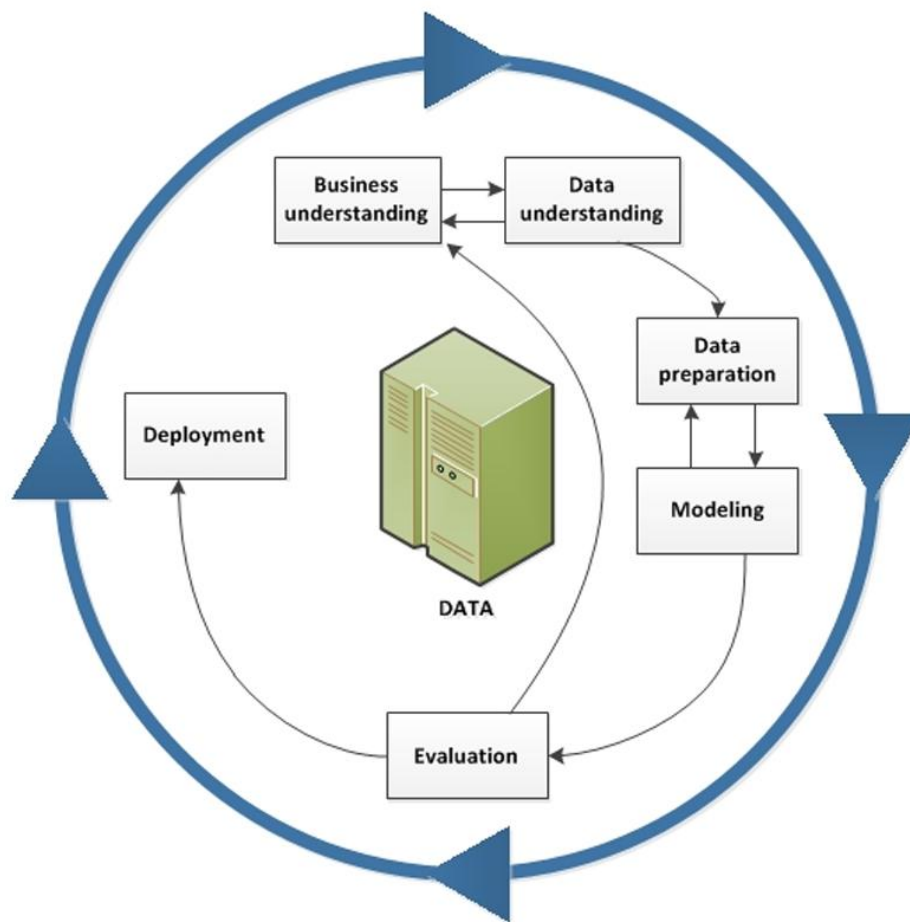


Figura 14: Metodología CRISP – DM para minería de datos [6]

3.1. Fases

A continuación, se muestra la metodología CRISP – DM [25] aplicada específicamente a la predicción del deterioro de la cartera crediticia, ajustando cada fase y tareas en relación al proyecto, con el fin de abordar los desafíos asociados de integrar variables externas como el cambio climático y el conflicto armado.

3.2. Entendimiento del negocio

Se investigó que el Banco cuenta con un indicador de cartera total vencido del 9.51% que corresponde a \$281.621 millones. La cartera de la banca **agropecuaria** de la regional occidente cuenta con indicador de cartera vencido del 12.16% que corresponde a \$223.217 millones y en con corte a julio del 2025, pero su explicación frente a este fenómeno, no estaba muy claro, con modelos clásicos que no absorbían la realidad de morosidad de este tipo de cartera, que no solo estaba relacionada con el cliente y sus impagos, sino que había un trasfondo mucho más complejo no se media teniendo los datos, siendo así, se definió el problema de investigación a resolver y su

relevancia para la cartera agropecuaria del Banco en la regional occidente. Se planteó la necesidad de evaluar diferentes modelos clásicos como de *Machine Learning* y encontrar patrones ocultos para poder exponerlos en un *dashboard* interactivo que permita tomar decisiones de manera informadas.

3.3. Entendimiento de los datos

3.3.1. Análisis exploratorio de los datos

El análisis exploratorio de datos (conocido de aquí en adelante como EDA) se debe aplicar antes de realizar cualquier modelo de predicción, esto es una de las fases más esenciales y que generalmente conlleva más tiempo en la ciencia de datos. En este apartado se realiza un EDA sobre las variables relacionadas con la cartera crediticia, el conflicto armado y el cambio climático. Esta fase incluye, revisar la cantidad de datos, la relación entre variables, detección de datos nulos o atípicos, revisar duplicidad de los datos, realizar estadística descriptiva y analizar su distribución, con el fin de asegurar la calidad de la información, descubrir patrones relevantes y obtener una base de datos robusta y eficiente. Para más información sobre el código, revisar el **Anexo 2**.

Es importante tener en cuenta que la entidad financiera cuenta con acuerdos de transferencia de datos con la Unidad para las Víctimas, la UNGRD y el IDEAM, cuya integración al dataset interno se realizó mediante llaves de cruce específicas para cada fuente, con fecha de corte unificada al 31/07/2025.

Con la Unidad para las Víctimas, la llave de integración es el número de identificación del cliente (CC/NIT). Esta entidad envía el reporte de quienes han sido reconocidos como víctimas del conflicto armado, indicando si el hecho victimizante ocurrió antes o después del desembolso del crédito, junto con la clasificación del hecho correspondiente. Los clientes del dataset interno sin coincidencia en esta fuente fueron conservados en el análisis, imputando la categoría "No reporta" en la variable `hecho_victimizante`, dado que la ausencia de registro no equivale necesariamente a no haber sido víctima, sino a no estar formalmente registrado ante la Unidad.

Con el IDEAM y la UNGRD, la llave de integración es la ciudad de inversión del crédito, complementada con las coordenadas geográficas (longitud y latitud) cuando el nivel municipal no permitía una coincidencia unívoca. A través de esta fuente se incorporan las afectaciones climáticas registradas en cada municipio por diversas causas: fenómeno de La Niña, deslizamientos, sequías, entre otras, las cuales alimentan la variable `Tipo_Factor_Externo` del dataset. Los municipios del dataset interno sin coincidencia en los registros de afectación climática fueron imputados con la categoría "Económico", correspondiente a la categoría de referencia del modelo, dado que la ausencia de reporte de afectación climática se interpreta como ausencia del evento en el período analizado.

La validación de consistencia entre fuentes se realizó verificando que el total de registros tras cada cruce no presentara una pérdida superior al 1% del dataset base, garantizando la integridad del dataset final de 55.433 observaciones utilizado en el modelado.

3.3.1.1 Descarga y descripción del *dataset*

Se utilizaron diferentes librerías, como Pandas, Numpy, Seaborn, entre otras para realizar el EDA y la manipulación de la base de datos.

La **Figura 15** nos muestra la cantidad de filas y columnas que posee el *dataset*, además el nombre de las columnas.

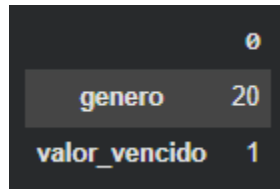
```
Cantidad de filas: 55474
Cantidad de columnas: 35
```

```
zonal
nombre_oficina
id
dias_vencido
rango
fecha_prox_ven
saldo
Banca_cliente
tipo_cartera
genero
ciudad_inversion
fecha_inicio_op
monto_desembolso
calificacion_obligacion
mo_modalidad
valor_vencido
saldo_cuota_prox_venc
etapa
estado_obligacion
cuotas_pagadas
cuotas_pactadas
cero_cuotas_pagadas
reestructurada
num_reestructuraciones
hecho_victimizante
fecha_ven_op
destino_agrupado
tipo_de_producto
tipo_productor_pmg
tipo_identificacion
tasa_referencial
COMPARATIVO_OCURRENCIA_HECHO
Factor Externo
Saldo Potencial
Tipo Factor Externo
```

Figura 15: Información sobre el *dataset*

Una vez obtenidos los datos, se procede a entender los datos, su significado e importancia para el modelo, para más información del significado de las variables revisar el **Anexo 1**.

Se empieza revisando los valores nulos de cada atributo, que se pueden ver la **Figura 16**



	0
genero	20
valor_vencido	1

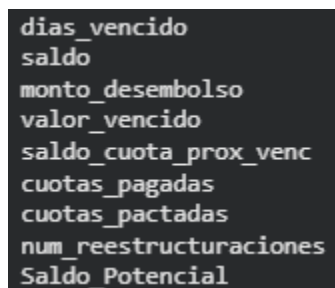
Figura 16: Variables con valores nulos

Estos registros que aparecen en la **Figura 16** que tienen valores nulos de los cuales fueron excluidos 21 del análisis, debido a que representan menos del 1% del total de observaciones. Su baja proporción implica una contribución estadísticamente insignificante al estudio y, por el contrario, podría generar distorsiones en los resultados o afectar la estabilidad de los métodos utilizados, por lo cual se opta por eliminar estas variables.

3.3.1.2 Estadísticas descriptiva – análisis univariado y bivariado

Una vez comprendido que datos que contiene el *dataset* es fundamental desarrollar las estadísticas descriptivas correspondientes al análisis univariado y bivariado, considerando tanto variables numéricas como categóricas. Este procedimiento permite caracterizar la distribución individual de cada variable y examinar sus correspondencias iniciales.

Con respecto a las variables numéricas, en la **Figura 17** podemos observar las variables que están incluidas en esta categoría y el resultado de la variable **Monto Desembolsado**:



```
dias_vencido  
saldo  
monto_desembolso  
valor_vencido  
saldo_cuota_prox_venc  
cuotas_pagadas  
cuotas_pactadas  
num_reestructuraciones  
Saldo_Potencial
```


Figura 17: Variables numéricas – descripción variable monto desembolsado

Las métricas (millones de pesos) hacen referencia a:

- **Count:** Cantidad total de registros
- **Mean:** Promedio de los valores de cada feature

```
=====
ANÁLISIS UNIVARIADO: monto_desembolso
=====

Resumen estadístico:
count      55454.00
mean      11713960.64
std       10696793.09
min        480000.00
25%        6000000.00
50%       10000000.00
75%       15000000.00
max       500000000.00
Name: monto_desembolso, dtype: float64

Sesgo (skew): 11.36331992465185
Curtosis (kurtosis): 306.79000971962904
```

- **Std:** Es la desviación estándar, que mide la dispersión o cuan alejados están los datos con respecto a la media
- **Min:** Valor mas pequeño del conjunto de datos
- **25% (1er cuartil):** Indica el valor por debajo en el que se encuentra el 25% de los datos
- **50% (Mediana):** Indica el valor central del conjunto de datos
- **75% (3er cuartil):** Indica el valor por debajo en el que se encuentra el 75% de los datos
- **Max:** Valor mas grande del conjunto de datos
- **Sesgo:** Indica cuan sesgada es la distribución con respecto a su media
- **Curtosis:** Indica que tan pesadas o ligeras son las colas de la distribución de los datos comparada con a su media

La variable **Monto Desembolsado** muestra una **media de aproximadamente 11,7 millones de pesos**, con una dispersión considerable (**desviación estándar de 10,7 millones**). Los valores típicos se concentran entre **6 millones (Q1)** y **15 millones (Q3)**, mientras que el valor máximo alcanza los **500 millones**, evidenciando la existencia de desembolsos extremadamente altos. Es importante tener en cuenta que los valores analizados corresponden a la cartera real comprendida al 31/07/2025.

Existe un sesgo positivo elevado (**skew = 11,36**) indica que la distribución está fuertemente inclinada hacia la derecha, es decir, la mayoría de los desembolsos son relativamente bajos

comparados con un pequeño grupo de desembolsos muy grandes. Esto se confirma con la curtosis demasiado alta (*kurtosis* = 306,79), lo que sugiere una distribución con colas pesadas y la presencia de múltiples valores atípicos significativos.

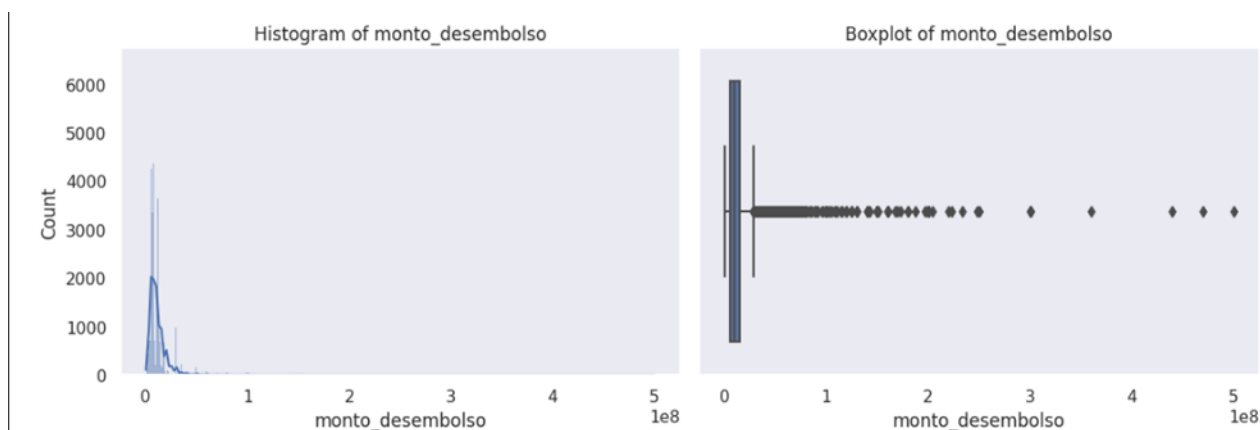


Figura 18: Histograma y Boxplot de la variable Monto de Desembolso

El **histograma** evidencia una acumulación en el rango inferior y la presencia de valores aislados muy grandes. El **boxplot** confirma esta situación: el rango intercuartílico es muy estrecho en comparación con los valores máximos, y se observa una gran cantidad de *outliers*, lo que refleja una **alta asimetría positiva** y **colas pesadas**. La distribución del **Monto de Desembolso** está concentrada en valores bajos y con pocos montos extremadamente altos, lo que genera una cola larga hacia la derecha.

Los resultados en conjunto dan un acercamiento al comportamiento de la variable **Monto Desembolsado** está fuertemente influenciado por grandes montos de préstamos. Para inspeccionar el resto de las variables numéricas, revisar **anexo 2**.

Con respecto a las variables categóricas, la **Figura 19** podemos observar las variables que están incluidas en esta categoría y el resultado de la variable **Factor Externo** mediante una tabla de frecuencia.

```
[ 'zonal',
  'rango',
  'Banca_cliente',
  'genero',
  'calificacion_obligacion',
  'mo_modalidad',
  'etapa',
  'estado_obligacion',
  'cero_cuotas_pagadas',
  'reestructurada',
  'hecho_victimizante',
  'destino_agrupado',
  'tipo_de_producto',
  'tipo_productor_pmg',
  'tipo_identificacion',
  'tasa_referencial',
  'COMPARATIVO_OCURRENCIA_HECHO',
  'Factor Externo',
  'Tipo Factor Externo']
```

Figura 19: Variable de la base de Datos

Tabla de Frecuencias — Variable Tipo de Factor Externo				
#	Tipo Factor Externo	Frecuencia	Porcentaje (%)	Porcentaje Acumulado (%)
0	Ola Invernal	39.729	0,716432	0,716432
1	Económico	15.078	0,271901	0,988333
2	Sequía	441	0,007953	0,996285
3	Factores Naturales - Volcán	195	0,003516	0,999802
4	Orden Público	10	0,000180	0,999982
5	Sequía y Heladas	1	0,000018	1,000000
	TOTAL	55.454	1,000000	

Tabla 8: Tabla de frecuencia variable tipo de factores externos

La **Tabla 8** muestra una tabla de frecuencia con los distintos tipos de factores externos (cambio climático) asociados a los eventos analizados. Se observa una clara preponderancia de la categoría “Ola Invernal”, que concentra el 71.64% de los registros, seguida de “Económico”, con un 27.19%. Estas dos categorías representan conjuntamente más del 98% de los casos, evidenciando que son los factores externos más relevantes en el conjunto de datos.

La **entropía** permite evaluar de manera general la diversidad en la distribución de una variable categórica. En este estudio, la variable Tipo de Factor Externo, asociada a eventos de cambio climático, presenta una **entropía de 0.65**, lo que indica una heterogeneidad media-alta

(balanceadas) entre sus categorías. Este indica que los distintos factores climáticos se encuentran relativamente bien distribuidos, aportando información relevante para el análisis del contexto externo que incide sobre la cartera de créditos agropecuarios.

Continuando con las variables categóricas, la **Tabla 9** muestra la tabla de frecuencia para la variable **hecho victimizante** asociada al tipo de conflicto armado reportado por el cliente, donde se observa que el 71% de los clientes no reporta haber sido víctima del conflicto armado, el 25% reporta que fue víctima de desplazamientos y despojos forzados y el 2% reporta que sufrieron homicidio o desaparición, siendo estos casos los más relevantes en el conjunto de datos.

Tabla de Frecuencias — Variable Hecho Victimizante				
#	Tipo Factor Externo	Frecuencia	Porcentaje (%)	Porcentaje Acumulado (%)
0	No reporta	39.080	0,7050	0,7050
1	Desplazamiento y despojo	13.803	0,2490	0,9540
2	Homicidio y desaparición	1.269	0,0229	0,9769
3	Atentados y amenazas	610	0,0110	0,9879
4	Otros	671	0,0121	1,0000
	TOTAL	55.433	1,0000	

Tabla 9: Tabla de frecuencia variable hecho victimizante

Evaluando la **entropía** de la variable tiene un **valor de 0.78**, lo que indica un alto balanceo entre las categorías de la variable, ósea, se encuentra relativamente bien distribuidos que aportan información relevante. Para inspeccionar el resto de las variables categóricas, tablas de frecuencia, entropía y matriz de correlación revisar **anexo 2**.

3.3.1.3 PCA mixto

Para abordar este punto, se emplearon metodologías de análisis multivariado destinadas a la selección y descripción de variables críticas. Se implementó el Análisis de Componentes Principales (PCA), priorizando aquellas dimensiones que capturen un umbral mínimo del 75% de la varianza acumulada. Este proceso se desglosará en tres etapas: un estudio de componentes para datos cuantitativos, un análisis de correspondencias para datos cualitativos y una integración final mediante PCA mixto.

3.3.1.4 MCA

El análisis de correspondencias múltiples por sus siglas en inglés (MCA) es una técnica estadística exploratoria utilizada para analizar y visualizar la relación entre variables categóricas al mismo tiempo. Así mismo permite la reducción de dimensionalidades, identificar patrones o perfiles en

los datos, entre otras funcionalidades.

El uso de este análisis fue fundamental para gestionar la alta dimensionalidad de la base de datos, compuesta principalmente por variables categóricas sobre el perfil del cliente y las condiciones del crédito. Este método estadístico permitió identificar las asociaciones más significativas entre las categorías y reducir el ruido, seleccionando variables como el tipo de producto, la tasa referencial y el destino de la inversión que capturaban la mayor parte de la inercia (variabilidad) del modelo. La eliminación de variables se basó en criterios de redundancia y bajo aporte explicativo, asegurando que el modelo predictivo se concentrara en los patrones con mayor capacidad de diferenciar el comportamiento de la cartera en la Regional Occidente.

Complementando el rigor estadístico, se aplicó un criterio de conocimiento del negocio para reintegrar y ajustar variables que, aunque presentan una naturaleza compleja, son determinantes en el contexto agropecuario colombiano. Variables como el "Hecho Victimizante", el "Comparativo de Ocurrencia", "Factores Externos Climáticos" y "monto desembolsado" fueron priorizadas debido a su relevancia crítica para entender el impacto del conflicto armado y el cambio climático en la mora. Esta selección híbrida garantiza que el modelo no solo sea matemáticamente robusto, sino que posea una alta interpretabilidad sectorial [12], permitiendo que los resultados reflejen fielmente la realidad socio-ambiental de los productores y su capacidad de pago frente a choques externos.

4. Modelado

En esta fase, correspondiente a la cuarta etapa de la metodología *CRISP-DM*, se implementaron cuatro familias de modelos predictivos para estimar el comportamiento de mora en la cartera agropecuaria: regresión logística binomial (glm), regresión logística ordinal (polr), árbol de decisión (rpart) y Extreme Gradient *Boosting* (XGBoost) . Cada familia fue entrenada en dos versiones: sin ponderación y con ponderación de clases, resultando en ocho modelos evaluados en total. Una vez concluida la fase de análisis exploratorio y selección de variables mediante MCA, se procedió a la preparación final del *dataset*. Este proceso comprendió cuatro etapas: (i) definición y codificación de la variable objetivo, (ii) tratamiento del desbalance de clases, (iii) selección y transformación de variables predictoras, y (iv) partición del dataset en conjuntos de entrenamiento y prueba.

Es importante precisar que los modelos implementados en esta fase son de naturaleza predictiva y asociativa, no causal. Su propósito es identificar patrones estadísticamente significativos entre las variables crediticias, climáticas y de conflicto armado con la probabilidad de mora, permitiendo la segmentación del riesgo, la generación de alertas tempranas y la priorización de intervenciones. En consecuencia, los resultados no permiten afirmar que el cambio climático o el conflicto armado causan directamente el deterioro de la cartera, sino que existe una asociación estadística significativa entre dichos factores y el comportamiento de mora en el período analizado. Una inferencia causal requeriría diseños experimentales o cuasi-experimentales con identificación exógena de la variación en las variables de interés, lo cual excede el alcance del presente proyecto

aplicado.

Respecto a las variables externas utilizadas en el modelado, es pertinente aclarar que tanto Tipo_Factor_Externo como hecho_victimizante provienen de registros internos del Banco, los cuales se alimentan de los cruces periódicos realizados con información oficial del IDEAM, la UNGRD y la Unidad para las Víctimas, tal como se describió en la sección 3.3.1. No se utilizaron datos directos del CERAC ni del IDEAM en formato de series de tiempo geoespaciales independientes; en cambio, dichas fuentes externas fueron incorporadas al dataset institucional mediante los acuerdos de transferencia de datos descritos previamente. Esta distinción es relevante para interpretar correctamente el alcance de las variables climáticas y de conflicto en el modelo: reflejan la clasificación institucional del Banco sobre las afectaciones de sus clientes, y no una medición directa e independiente de los fenómenos climáticos o del conflicto armado en el territorio.

4.1 Definición de la variable objetivo

Se definieron dos formulaciones de la variable objetivo, correspondientes a las dos estrategias de modelado implementadas:

Variable binaria (mora sí/no): Se construyó una variable dicotómica que clasifica cada obligación como "Al día" (0 días de vencimiento) o "En mora" (1 o más días de vencimiento). Esta formulación sustenta el modelo de regresión logística binomial. De un total de 55.454 obligaciones, 48.263 se encontraban al día (87.0%) y 7.191 en mora (13.0%), lo que evidencia un desbalance de clases de aproximadamente 6.7 es decir, por cada obligación en mora existen 6.7 al día.

Variable ordinal (severidad de mora): Se definió una variable ordinal con tres categorías, construida exclusivamente sobre las obligaciones en mora ($\text{dias_vencido} > 0$): mora temprana (1–30 días), mora media (31–90 días) y mora severa (más de 90 días). El modelo ordinal opera sobre este subconjunto ($n = 7.191$), dado que su objetivo es estratificar la intensidad del deterioro, no predecir si hay mora. La distribución observada fue: 977 obligaciones en rango 1–30 días (13.6%), 2.671 en rango 31–90 días (37.1%) y 3.543 con más de 90 días (49.3%), con un desbalance de 3.63:1 entre la clase mayoritaria (mora severa, 91+ días) y la minoritaria (mora temprana, 1–30 días), es decir, por cada obligación con mora temprana existen aproximadamente 3.63 con mora severa. Este desbalance justifica la ponderación diferencial de clases aplicada en el modelo ordinal

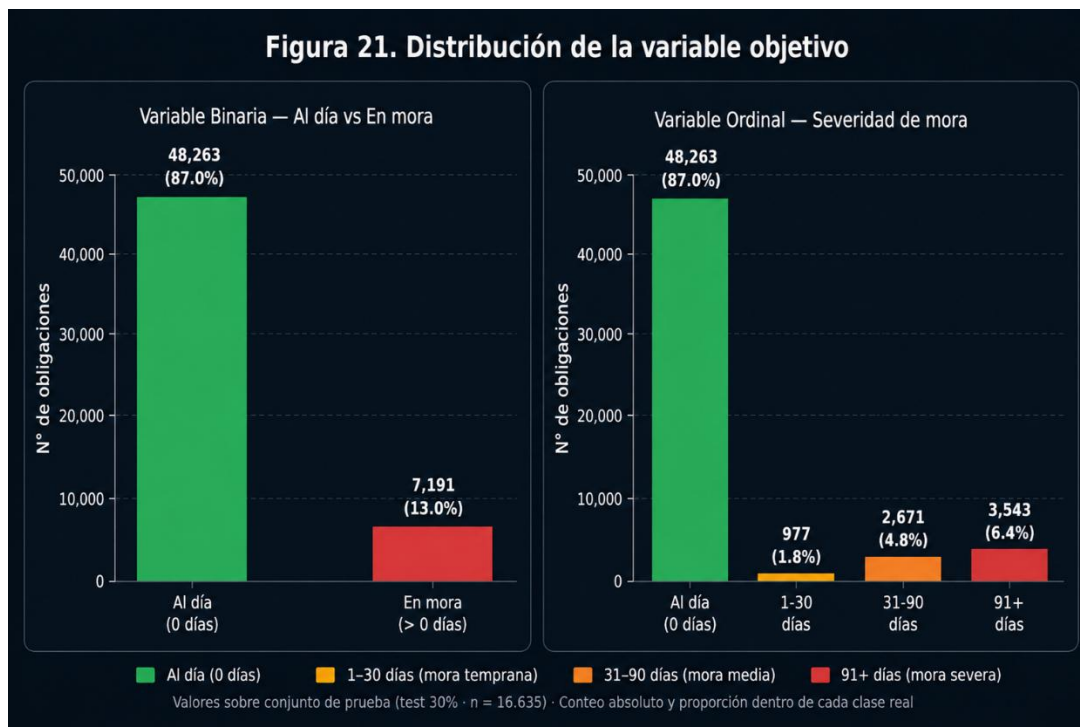


Figura 20: Distribución de la variable objetivo: mora binaria (izquierda) y severidad ordinal (derecha).

4.1.1 Tratamiento del desbalance de clases

Dado el marcado desbalance de clases identificado en la variable objetivo (87.03% al día vs. 12.97% en mora), se implementó una estrategia de ponderación de observaciones (*class weights*) en lugar de técnicas de remuestreo como *SMOTE* o *undersampling*. Esta decisión responde a dos consideraciones: primero, el *dataset* presenta un número absoluto suficiente de observaciones en la clase minoritaria para sostener el entrenamiento; segundo, la ponderación conserva la distribución original de los datos sin distorsionar artificialmente la estructura poblacional de la cartera.

Para el modelo binomial, los pesos se calcularon fijando una proporción objetivo de 55% (Al Día) / 45% (En Mora), resultando en pesos de 0.6320 y 3.4701 respectivamente ($\text{weights_class} = \text{prop_objetivo} / \text{prop_real}$). Para el modelo ordinal, la proporción objetivo fue 30% (1–30 días) / 35% (31–90 días) / 35% (91+ días), compensando el severo desbalance de la clase de mora temprana (13.6% en el subconjunto de mora). En ambos casos, el vector de pesos se asignó por observación según su clase real, sin modificar la distribución del conjunto de prueba.

4.1.2 Variables predictoras seleccionadas

A partir del Análisis de Correspondencias Múltiples (MCA) y el conocimiento del negocio documentado en la sección 3.3.1.4, se seleccionó un conjunto común de variables (`var_sele`) como punto de partida. Las fórmulas finales difieren entre modelos: el binomial (`glm`) excluye `hecho_victimizante` de la fórmula de estimación, mientras que el ordinal (`polr`) la incluye. Las variables `días_vencido` y `rango` se usan únicamente para construir la variable objetivo y se eliminan explícitamente antes del XGBoost por constituir *data leakage*. Las variables numéricas `cuotas_pactadas`, `monto_desembolso` y `saldo` fueron estandarizadas (media 0, desviación estándar 1) usando `preProcess()` de `caret`, aprendido solo sobre el conjunto de entrenamiento.

A continuación se detallan las 13 variables predictoras utilizadas en los modelos binomiales (`glm`, `rpart`, `XGBoost`) y las 12 del modelo ordinal (`polr`):

- **Tipo Factor Externo (cambio climático):** Ola Invernal, Económico, Sequía, Factores Naturales–Volcán, Orden Público.
- **Hecho victimizante (conflicto armado):** No reporta, Desplazamiento y despojo, Homicidio/desaparición, Atentados/terrorismo, Otros.
- **Monto desembolso:** variable numérica continua, transformada mediante logaritmo natural para corregir la alta asimetría positiva ($skew = 11.36$).
- **Tipo de producto:** Inversión Ordinaria FINAGRO, Inversión víctima conflicto armado, Capital de trabajo, entre otros.
- **Tasa referencial:** IBR Semestral, DTF Efectivo Anual, IBR Trimestral, TASA Cero, entre otras.
- **Tipo de productor:** Pequeño Productor, Microfinanzas, No Agropecuario, Mediano Productor, Empresarial Cadena Agropecuaria, Grandes Productores Finagro.
- **Destino agrupado:** Principales categorías: Café, Ganadería y Pecuaria, Microfinanzas, Agricultura y cultivos, Consumo, Caña panelera, Infraestructura, entre otros. Categoría de referencia: Pecuaria y pesca.
- **Cuotas pactadas:** Número total de cuotas pactadas al momento del desembolso del crédito.
- **Comparativo de ocurrencia del hecho:** variable que indica si el cliente fue afectado antes o durante el período de análisis (V-D, D-V, No reporta).
- **Género del titular:** Masculino, Femenino.
- **Banca cliente:** Agropecuaria, personal, microfinanzas, oficial, empresarial y pyme consolidada.
- **Mo_Modalidad:** Tipo de Amortización (mes vencido, trimestre vencido, semestre vencido, año vencido) entre otras. En el subconjunto de mora los niveles son AV, MV, SV, TV.
- **Departamento de Inversión:** Variable categórica (`Departamento_Inversion`). Niveles con menos de 30 observaciones agrupados en “Other”. Departamentos retenidos:

Cauca, Nariño, Valle del Cauca, Other (incluye Chocó y otros minoritarios).

- **Saldo:** Saldo capital adeudado por el cliente a corte 31/07/2025
- **Rango:** clasificación de los días vencido.
- **Días Vencido:** número días de mora de la obligación.

Respecto a la temporalidad de la variable Saldo, es importante precisar que tanto esta variable como la variable objetivo (días_vencido) se evalúan al mismo corte transversal del 31/07/2025. El saldo capital adeudado en dicha fecha refleja el stock de deuda vigente en el momento del análisis y no incorpora información generada con posterioridad al evento de mora, dado que la mora y el saldo son medidos simultáneamente en el mismo corte. En ese sentido, el saldo no constituye una variable con fuga de información en sentido estricto, pues no anticipa ni es consecuencia directa de la clasificación binaria de mora utilizada como variable objetivo, la cual se define únicamente por los días de vencimiento registrados a esa misma fecha.

Las variables categóricas fueron convertidas a tipo factor en R, lo que permite que los modelos generen automáticamente las variables indicadoras internas, tomando como categoría de referencia la modalidad con mayor frecuencia en cada variable. Para la función `polr()`, las variables de los niveles presentes en *test* pero no en *train* se alinean mediante una función ad hoc (`alinear_niveles`) que convierte esas observaciones en NA, preservando el tamaño real del conjunto de prueba y garantizando consistencia metodológica en la evaluación.

4.1.3 Partición *train/test*

El *dataset* fue dividido en conjunto de entrenamiento (70%) y conjunto de prueba (30%), mediante muestreo estratificado con `createDataPartition()` para garantizar que la proporción de obligaciones en mora se conservara en ambas particiones (Al Día: 87.03%; En Mora: 12.97%). La partición estratificada es especialmente crítica en *datasets* con desbalance de clases, ya que una partición aleatoria simple podría generar subconjuntos con distribuciones significativamente diferentes. El conjunto de entrenamiento quedó conformado por 38.819 observaciones y el conjunto de prueba por 16.635, asegurando una evaluación robusta del poder predictivo de los modelos sobre datos no vistos.

4.1.4 Modelado — Regresión Logística Binomial

El primer enfoque de modelado se basó en la regresión logística binomial (logit binario), una técnica estadística ampliamente utilizada en el análisis de riesgo crediticio por su interpretabilidad y robustez ante supuestos de linealidad en la función logit. Este modelo estima la probabilidad de

que una obligación entre en mora (variable binaria: Al Día vs. En Mora) a partir de las 13 variables predictoras descritas en la sección 3.3.2.3.

El modelo fue implementado en R mediante la función `glm()` con familia binomial y función de enlace logit, evaluado sobre el conjunto de prueba ($n = 16.635$). Se estimaron dos versiones: sin ponderación de clases (Sin pesos, umbral Youden = 0.1479) y con ponderación de clases (Ponderado, proporción objetivo 55/45, umbral Youden = 0.4927), para evaluar el impacto de la corrección del desbalance sobre las métricas de desempeño.

Modelo Binomial sin peso: Este modelo obtuvo un AUC-ROC de 0.6493, Accuracy de 0.6836, *Sensitivity* de 0.5155, *Specificity* de 0.7087, *Balanced Accuracy* de 0.6121, Kappa de 0.1379 y F1 de 0.2970. La matriz de confusión muestra 10.260 verdaderos negativos, 1.045 falsos negativos, 4.218 falsos positivos y 1.112 verdaderos positivos. Estos resultados evidencian que, sin ponderación, el modelo sesga sus predicciones hacia la clase mayoritaria (Al Día), sacrificando la capacidad de detección de mora.

Modelo Binomial ponderado: Al incorporar la ponderación de clases (proporción objetivo 55/45, umbral Youden = 0.4927), el AUC-ROC se mantuvo en 0.6493, con Accuracy de 0.7000, *Sensitivity* de 0.4975, *Specificity* de 0.7302, *Balanced Accuracy* de 0.6138, Kappa de 0.1462 y F1 de 0.3007. La matriz registra 10.572 verdaderos negativos y 1.073 verdaderos positivos. Este *trade-off* — mayor *Specificity* a costa de algo de *Sensitivity* — es esperado y deseable en gestión de riesgo crediticio, donde el costo de no detectar un cliente en mora (falso negativo) supera al de clasificar erróneamente uno al día (falso positivo).

4.2 Modelado — Árbol de Decisión (`rpart`)

Como tercer enfoque de modelado se implementó un árbol de decisión mediante la función `rpart()` con parámetro de complejidad $cp = 0.01$, estimado en dos versiones: sin ponderación y con ponderación de clases. Este método genera reglas de clasificación interpretables y no requiere supuestos de distribución, lo que lo convierte en una línea de referencia natural para evaluar el valor añadido por los modelos más complejos.

Árbol sin pesos: Obtuvo un AUC-ROC de 0.5550 con Accuracy del 87.13%, umbral Youden de 0.1546, *Sensitivity* de 0.1261, *Specificity* de 0.9823, Kappa de 0.1598 y F1 de 0.2026. El elevado Accuracy refleja simplemente el sesgo hacia la clase mayoritaria (al día), con una capacidad discriminatoria próxima al azar ($AUC \approx 0.56$).

Árbol ponderado: La incorporación de pesos mejoró la *Sensitivity* de 0.1261 a 0.2707, con AUC de 0.5779, Accuracy de 0.8034, *Specificity* de 0.8828, Kappa de 0.1499 y F1 de 0.2632. Ambas versiones del árbol presentaron los AUC más bajos del portafolio (0.555 y 0.578), confirmando que la estructura del *dataset* requiere modelos de mayor capacidad de generalización como

XGBoost.

4.3 Modelado — XGBoost (Extreme Gradient Boosting)

El cuarto enfoque corresponde a XGBoost (Extreme Gradient *Boosting*), implementado mediante la función `xgb.train()` con parámetros `eta = 0.05`, `max_depth = 4`, `subsample = 0.6`, `colsample_bytree = 0.6`, `alpha = 1.0`, `lambda = 5.0`, `min_child_weight = 15`, `gamma = 0.3` y hasta 800 rondas con parada temprana a las 20 rondas sin mejora en AUC sobre el conjunto de prueba. Este es un modelo de ensamble que entrena secuencialmente árboles de decisión donde cada nuevo árbol corrige los errores del anterior. Con el fin de eliminar fuga de información (*data leakage*), las variables `dias_vencido` y `rango` fueron excluidas explícitamente de la matriz de *features* antes del entrenamiento.

XGBoost sin pesos: Obtuvo el segundo mejor AUC-ROC global de 0.7689, con Accuracy del 77.08%, umbral Youden de 0.1311, *Sensitivity* de 0.6291, *Specificity* de 0.7919, Kappa de 0.2931 y F1 de 0.4158. Estos resultados superan ampliamente a los modelos Logit y al árbol de decisión, confirmando la superior capacidad discriminatoria del *boosting* en presencia de variables categóricas con alta cardinalidad e interacciones no lineales.

XGBoost ponderado: Con la incorporación de pesos por observación en la *DMatrix* (proporción objetivo 55/45, umbral Youden de 0.4509), el modelo alcanzó el mejor resultado global del portafolio: AUC-ROC de 0.7921, Accuracy del 78.93%, *Sensitivity* de 0.6384, *Specificity* de 0.8118, Kappa de 0.3253 y F1 de 0.4400. El incremento en Kappa de 0.2931 a 0.3253 respecto a la versión sin pesos confirma una mejora sustancial en la concordancia real de las predicciones, más allá del azar, particularmente en la detección de clientes en mora.

La selección de los hiperparámetros finales del modelo XGBoost [31] se realizó mediante un proceso de búsqueda iterativa manual, evaluando el AUC-ROC sobre el conjunto de prueba como métrica de selección. Se priorizaron combinaciones que maximizaran el AUC-ROC con parada temprana a las 20 rondas sin mejora, lo que permitió controlar el sobreajuste sin necesidad de un grid search exhaustivo. Para el Árbol de Decisión, el parámetro de complejidad `cp = 0.01` fue seleccionado tras evaluar valores en el rango {0.001, 0.005, 0.01, 0.05}, eligiendo el valor que equilibró la profundidad del árbol con su capacidad de generalización sobre el conjunto de prueba.

4.4 Modelado — Regresión Logística Ordinal (severidad de mora)

El enfoque de modelado adoptó una perspectiva más granular del riesgo crediticio, mediante la regresión logística ordinal (modelo `polr`), que permite modelar la severidad de la mora como una variable con ordenamiento natural: el modelo opera con tres categorías: mora 1 - 30 días, 31 - 90 días y más de 90 días. A diferencia del modelo binario, que únicamente distingue entre clientes al día y en mora, el modelo ordinal permite identificar la intensidad del deterioro, información

esencial para diseñar estrategias de atención diferenciadas según el grado de afectación del cliente.

El modelo fue implementado en un entorno de desarrollo integrado de código abierto llamado R Studio mediante la función `polr()` del paquete MASS, con método logístico. Al igual que en el modelo binomial, se estimaron versiones con y sin ponderación de clases. Los pesos fueron calibrados de forma inversamente proporcional a la frecuencia de cada categoría ordinal, dado que la clase 1-30 días representa el 13.6% del subconjunto de mora. La suma de pesos por clase quedó equilibrada: 1.510,50 (1-30), 1.762,25 (31-90) y 1.762,25 (91>).

Modelo Ordinal sin pesos: Este modelo obtuvo un AUC macro de 0.5991, Accuracy de 0.4620. Por clase: *Sensitivity* 1-30 = 0.034, *Sensitivity* 31-90 = 0.237, *Sensitivity* 91> = 0.750. Las *Balanced Accuracy* fueron 0.513 (1-30), 0.464 (31-90) y 0.543 (91>). La matriz de confusión concentra predicciones en la clase 91>, con solo 10 de 293 casos 1-30 días correctamente identificados. La Accuracy inferior es propia de la clasificación en tres clases. Esta primera línea los 30 días fue de apenas 0.034, evidenciando la incapacidad del modelo para detectar la mora incipiente en ausencia de ponderación. La *Balanced Accuracy* media para las tres clases fue de 0.507, equivalente prácticamente al azar.

Modelo Ordinal ponderado: La incorporación de pesos (proporción objetivo 30/35/35 por clase) produjo mejoras sustanciales en la detección de mora temprana: la *Sensitivity* para la clase 1-30 días se incrementó de 0.034 a 0.614, la *Balanced Accuracy* para dicha clase pasó de 0.513 a 0.717, y el AUC macro mejoró de 0.5991 a 0.6608. Con una Accuracy del 43.37%. Estos resultados posicionan al Logit Ordinal Ponderado como el modelo con mayor capacidad para detectar deterioro temprano en la cartera agropecuaria dentro de la familia de modelos ordinales.

La **Figura 21** presenta la distribución de la severidad de mora desagregada por departamento y tipo de factor externo climático, revelando la heterogeneidad territorial del deterioro.

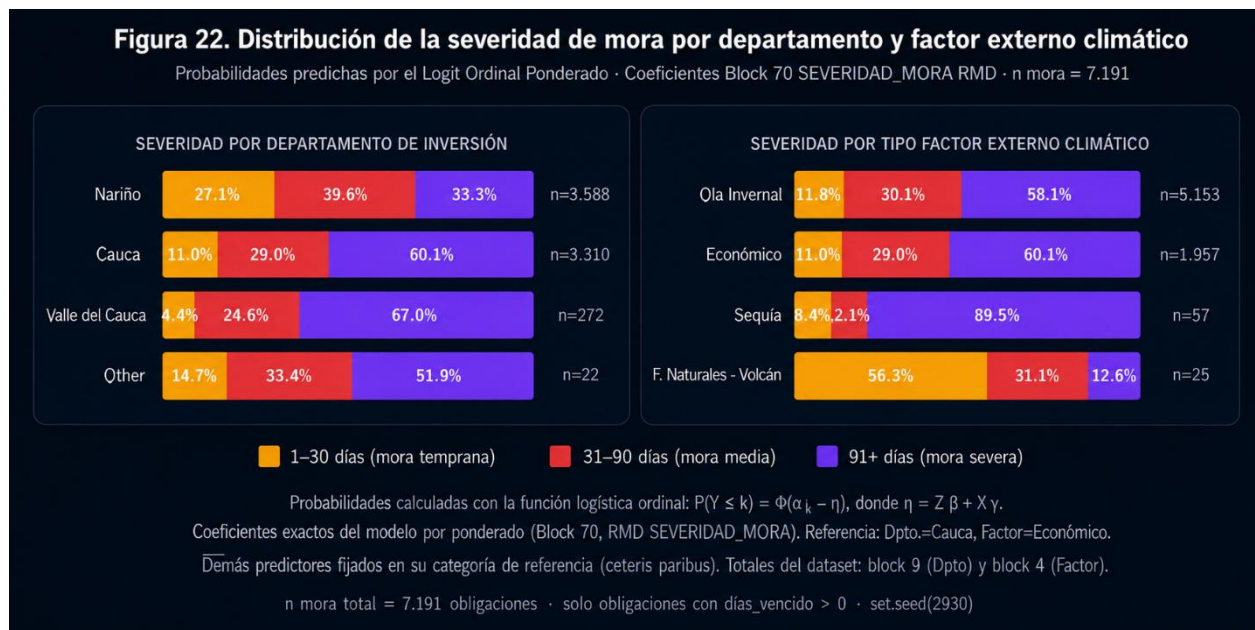


Figura 21: Distribución de la severidad de mora por departamento (izquierda) y tipo de factor externo climático (derecha).

La **Figura 21** presenta las probabilidades predichas de severidad de mora según el Logit Ordinal Ponderado. Por departamento: Nariño muestra la menor probabilidad de mora severa (91+) respecto a Cauca (coeficiente $\beta = -1.105$, $p < 0.001$), mientras que Valle del Cauca presenta mayor riesgo ordinal ($\beta = +0.300$). Por factor externo: Sequía se asocia con la mayor probabilidad de mora severa ($\beta = +1.732$, $p < 0.001$), mientras que Factores Naturales-Volcán muestra el mayor efecto protector contra mora severa ($\beta = -2.345$, $p < 0.001$). Ola Invernal no presenta diferencia significativa respecto al factor económico ($\beta = -0.081$, $p = 0.357$) en el modelo ordinal ponderado.

4.5 Comparación y selección del modelo

Con el propósito de identificar el modelo con mayor capacidad predictiva para los objetivos del presente estudio, se realizó una evaluación comparativa de los seis modelos binomiales estimados (Logit, Árbol de Decisión y XGBoost, cada uno con y sin ponderación de clases) [21], considerando un conjunto amplio de métricas de desempeño sobre el conjunto de prueba ($n = 16.635$).

Tabla 10: comparativo de métricas

Modelo	Tipo	AUC	Accuracy	Kappa	Bal. Acc.	Sensitivity	Specificity	Observación
Logit Binario sin pesos	Binario	0.6493	0.6836	0.1379	0.6121	0.5155	0.7087	Mejor F1 Logit
Logit Binario ponderado	Binario	0.6493	0.7000	0.1462	0.6138	0.4975	0.7302	Mayor Accuracy Logit
Árbol de Decisión sin pesos	Binario	0.5550	0.8713	0.1598	0.5542	0.1261	0.9823	Línea base árbol
Árbol de Decisión ponderado	Binario	0.5779	0.8034	0.1499	0.5768	0.2707	0.8828	Mayor Sensitivity árbol
XGBoost sin pesos	Binario	0.7689	0.7708	0.2931	0.7105	0.6291	0.7919	Segundo mejor AUC
XGBoost ponderado	Binario	0.7921	0.7893	0.3253	0.7251	0.6384	0.8118	Mejor AUC global ★

Tabla 10. Comparación de métricas de desempeño - Modelos de predicción de mora.

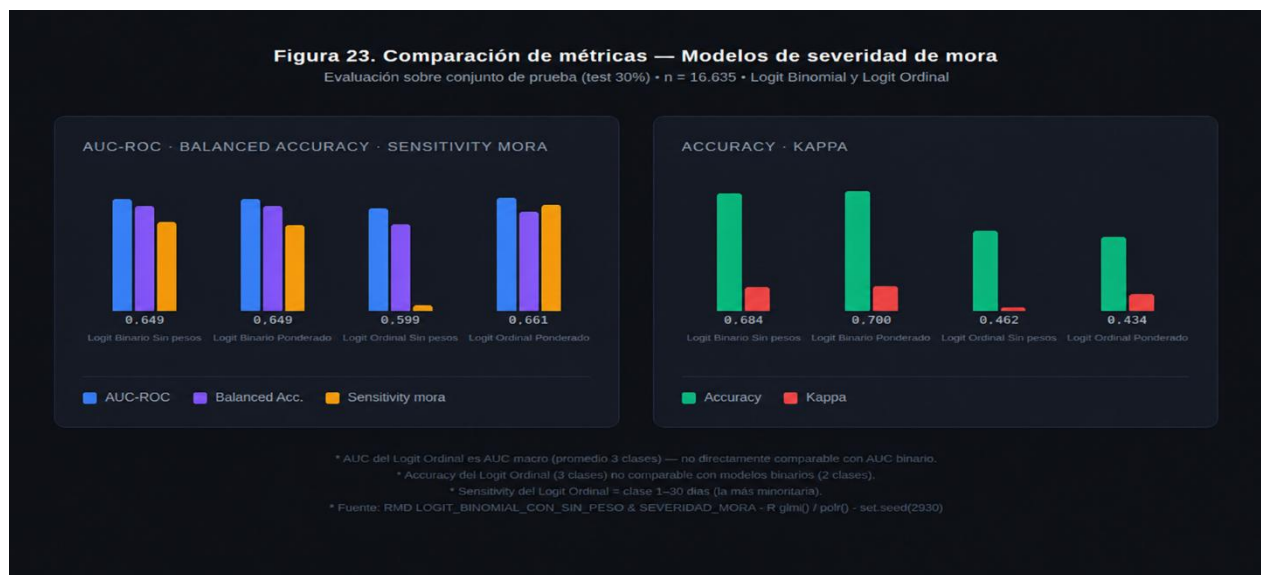


Figura 22: Comparación de métricas por modelo: AUC-ROC, *Balanced Accuracy*, *Sensitivity* (izquierda) y *Accuracy* (derecha).

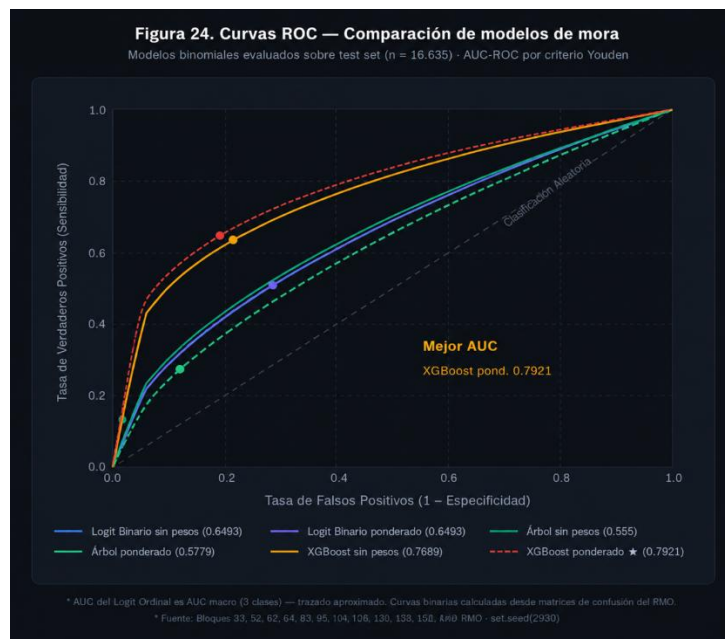


Figura 23: Curvas ROC comparadas para los seis modelos binomiales (Logit, Árbol de Decisión y XGBoost, con y sin ponderación).

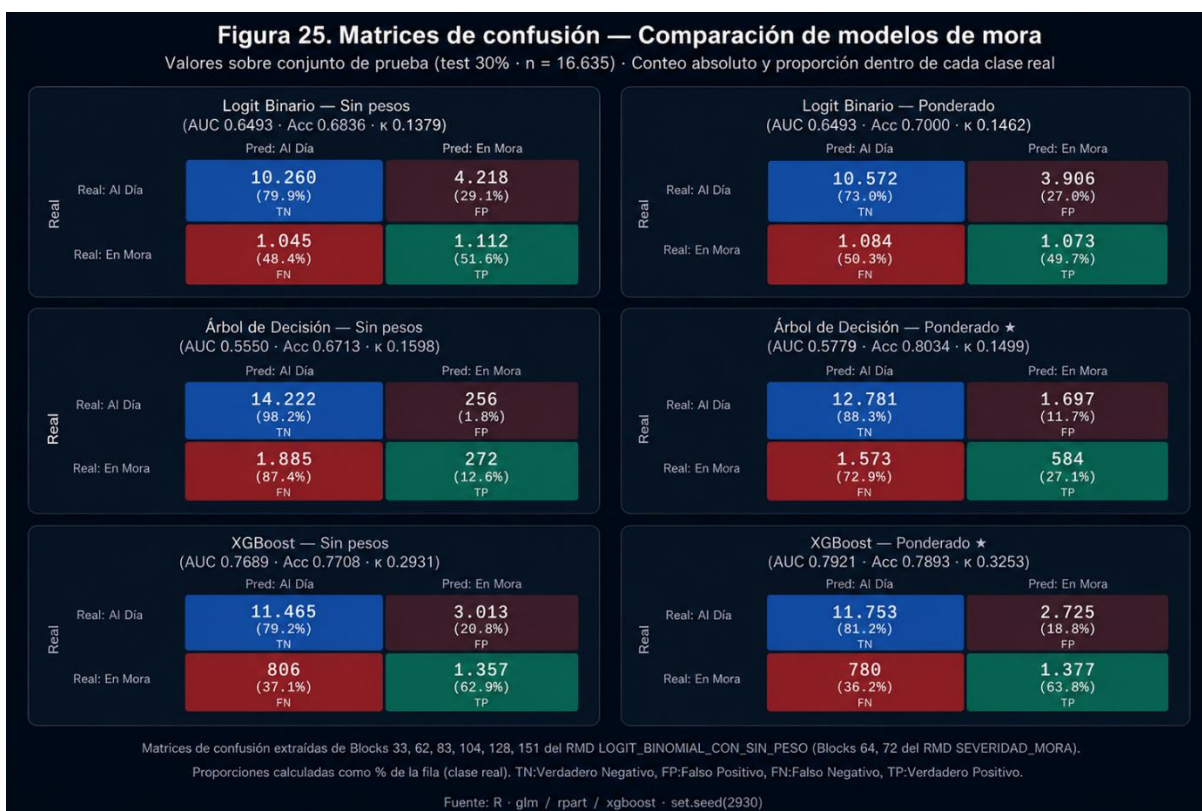


Figura 24: Matrices de confusión de los seis modelos binomiales: Logit, Árbol de Decisión y XGBoost

(con y sin ponderación).

Del análisis comparativo se desprenden las siguientes conclusiones metodológicas:

Para clasificación binaria masiva con máxima capacidad discriminatoria: el `XGBoost` ponderado es el modelo principal del portafolio ($AUC = 0.7921$, $Sensitivity = 0.6384$, $Specificity = 0.8118$, $F1 = 0.4400$), superando en ~ 14 puntos de AUC al mejor Logit Binomial y en ~ 21 puntos al Árbol de Decisión. Para alertas tempranas de mora incipiente (1–30 días): el Logit Ordinal Ponderado es el más adecuado, con AUC macro 0.6608 y $Sensitivity_{1-30} = 0.614$.

En consecuencia, se recomienda una arquitectura de tres modelos complementarios: (i) el `XGBoost` ponderado como motor principal de *scoring* binario, dado su superior capacidad discriminatoria ($AUC = 0.7921$); (ii) el Logit Ordinal Ponderado como herramienta de alerta temprana focalizada en clientes con primeras señales de deterioro ($Sensitivity_{1-30 \text{ días}} = 0.614$); y (iii) el Logit Binomial como instrumento de interpretación regulatoria y auditabilidad, dada su legibilidad en términos de odds ratios y coeficientes. Esta arquitectura complementaria maximiza la utilidad analítica en el contexto operativo de la entidad financiera, equilibrando precisión predictiva, detección temprana e interpretabilidad.

5. Evaluación y resultados del modelo predictivo

En este capítulo se presentan los resultados derivados de la implementación y evaluación de los ocho modelos predictivos de mora en la cartera agropecuaria de la Regional Occidente. El análisis está estructurado en cuatro componentes: (i) interpretación de los coeficientes y odds ratios del modelo Logit Binomial, (ii) análisis de la discriminación por severidad en el modelo ordinal, (iii) implicaciones para la gestión del riesgo crediticio agropecuario, y (iv) limitaciones del modelo. El modelo `XGBoost` ponderado, seleccionado como modelo principal, es discutido en su desempeño global dentro del análisis comparativo de la sección 3.3.5.

5.1 Interpretación de coeficientes

La regresión logística binomial permite cuantificar el efecto de cada variable predictora sobre la probabilidad de mora mediante el odds ratio (OR), definido como la razón de momios entre la categoría analizada y la categoría de referencia. Un OR mayor a 1 indica mayor probabilidad de mora, mientras que un OR menor a 1 indica efecto protector. Los OR se calcularon como $\exp(\beta)$ a partir de los coeficientes del modelo Logit Binomial Ponderado

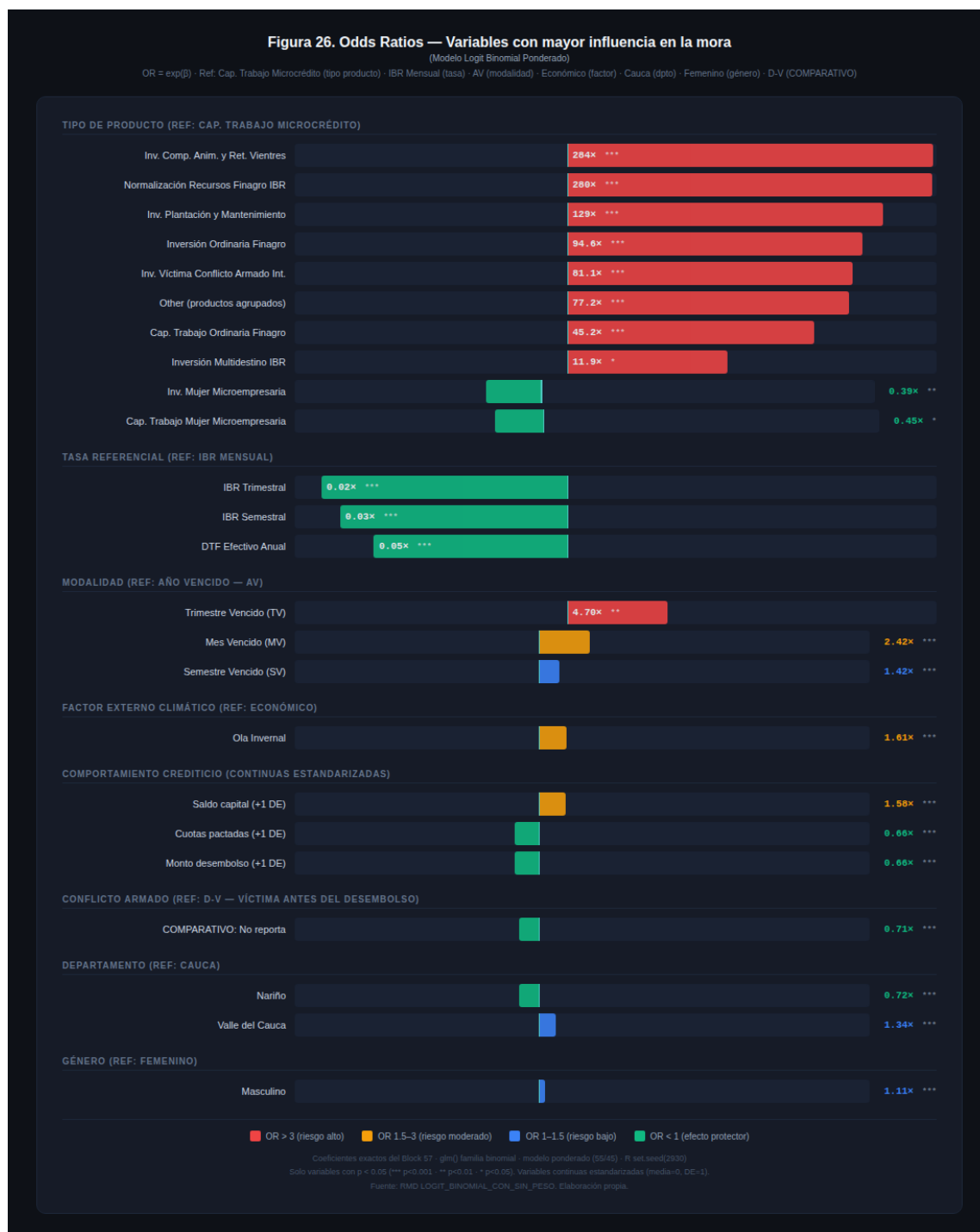


Figura 25: Odds Ratios del Modelo Logit Binomial - variables con mayor impacto sobre la probabilidad de mora. Elaboración propia.

La **Figura 25** presenta los odds ratios estimados para las variables con mayor influencia en la probabilidad de mora, donde se encuentran los siguientes hallazgos:

Género: Los hombres tienen una probabilidad de caer en mora un 11% mayor que las mujeres (OR = 1.11).

Factor externo climático: Las obligaciones afectadas por Ola Invernal tienen una probabilidad de caer en mora un 61% mayor que las afectadas por el factor Económico (OR = 1.61).

Cuotas pactadas: Por cada desviación estándar adicional en el número de cuotas pactadas, la probabilidad de mora se reduce un 34.3% (OR = 0.657), indicando que los créditos de mayor plazo tienen menor riesgo relativo.

Conflicto armado: Los clientes que no reportan afectación por conflicto tienen una probabilidad de caer en mora un 29% menor que quienes fueron víctimas antes del desembolso (OR = 0.714).

Departamento: Las obligaciones en Nariño tienen una probabilidad de caer en mora un 28% menor que las del Cauca (OR = 0.717). Por su parte, las obligaciones en Valle del Cauca tienen una probabilidad un 34% mayor que las del Cauca (OR = 1.341).

Tasa referencial: Los créditos referenciados a IBR Semestral tienen una probabilidad de caer en mora un 97% menor que los referenciados a IBR Mensual (OR = 0.030). Los referenciados a DTF E.A. tienen una probabilidad un 95% menor (OR = 0.051).

Tipo de producto: Los créditos de Inversión Ordinaria Finagro tienen una probabilidad de caer en mora significativamente mayor que los de Capital de Trabajo Microcrédito (OR = 94.6). Este valor extremo refleja en parte el reducido tamaño muestral de esta categoría y debe interpretarse con cautela.

5.2 Análisis de discriminación por severidad - Modelo Ordinal

El modelo logístico ordinal ponderado permite identificar no solo qué clientes tienen mayor riesgo de mora, sino también en qué nivel de severidad es probable que se ubiquen. Los umbrales estimados por el modelo (θ_1 y θ_2) definen los puntos de corte entre las categorías ordinales en la escala latente de riesgo.

Los resultados muestran que el Accuracy para la clase 1 - 30 días (mora temprana) alcanzó 0.717 en el modelo ponderado, frente a 0.513 en la versión sin peso. Esta mejora sustancial indica que la ponderación logró compensar efectivamente el déficit de información de la clase más infrecuente, sin degradar severamente la discriminación de las demás categorías.

Para la clase mayor a 91 días (mora severa), la *Sensitivity* fue de 0.488 en el modelo ponderado, reflejando una capacidad moderada de detección de mora avanzada. Este resultado es consistente

con la naturaleza de la regularización por pesos, que redistribuye el esfuerzo clasificatorio hacia las categorías más infrecuentes. La *Balanced Accuracy* promedio de las tres clases de mora mejoró de 0.507 (sin pesos) a 0.580 (ponderado), lo que representa una mejora del 14.4% en la capacidad de discriminación multiclase ajustada por el desbalance.

Desde una perspectiva de gestión, la capacidad del modelo ordinal de estratificar la cartera en tres niveles de mora permite diseñar respuestas diferenciadas: para clientes en mora 1 - 30 días, acciones preventivas inmediatas como llamadas de gestión y acuerdos de normalización voluntaria; para clientes en mora entre 31 - 90 días, visitas de campo y reestructuración; y para mora superior a 90 días, activación de garantías y procesos de recuperación.

5.3 Implicaciones para la gestión de riesgo crediticio agropecuario

Los modelos desarrollados permiten traducir los hallazgos analíticos en recomendaciones concretas para la gestión proactiva de la cartera agropecuaria en la Regional Occidente. Se identifican tres dimensiones de impacto:

Alertas tempranas basadas en variables externas: La integración del Factor Externo Climático y el Hecho Victimizante como predictores validados permite construir un sistema de alertas activado por eventos externos. Cuando el Banco registra una Ola Invernal o un evento de orden público en un municipio, el modelo puede reevaluar automáticamente el perfil de riesgo de las obligaciones vigentes en dicha zona, priorizando la gestión preventiva antes de que el deterioro se materialice en mora efectiva.

Segmentación territorial del riesgo: El análisis reveló heterogeneidad significativa entre departamentos. Las obligaciones en Valle del Cauca tienen una probabilidad de mora un 34% mayor que las del Cauca ($OR = 1.341$, $p < 0.001$), mientras que las de Nariño presentan una probabilidad un 28% menor ($OR = 0.717$, $p < 0.001$). Estas diferencias justifican políticas de colocación y seguimiento diferenciadas por territorio, con atención prioritaria en los municipios con mayor concentración de eventos de Ola Invernal y desplazamiento forzado, que el modelo identifica como factores de riesgo estadísticamente significativos.

Estrategias por ciclo de vida del crédito: El número de cuotas pactadas resultó ser uno de los predictores más significativos del modelo ($p < 0.001$). Por cada desviación estándar adicional en las cuotas pactadas, la probabilidad de mora se reduce un 34% ($OR = 0.657$), lo que indica que los créditos de mayor plazo relativo presentan menor riesgo de deterioro. Este hallazgo sugiere que las estrategias de acompañamiento deben concentrarse en los créditos de corto plazo y en los primeros períodos tras el desembolso, especialmente para clientes con hecho victimizante o afectados por factores externos climáticos.

5.4 Limitaciones del modelo

Los modelos desarrollados presentan un conjunto de limitaciones que deben considerarse en su interpretación y aplicación práctica:

- **Capacidad predictiva moderada:** los valores de AUC-ROC del portafolio oscilan entre 0.555 (Árbol sin pesos) y 0.792 (XGBoost ponderado). Los modelos Logit presentan AUC moderado (0.649), mientras XGBoost alcanza un nivel discriminatorio aceptable. Factores no observados como precios de mercado, condiciones individuales del productor y calidad de las garantías no están disponibles en el *dataset* y podrían mejorar el poder predictivo de los modelos más simples.
- **Desbalance de clases residual:** a pesar de los avances logrados con la ponderación, el Kappa máximo de los modelos Logit es 0.1462, lo que indica concordancia limitada más allá del azar. El XGBoost ponderado mejora este indicador a 0.3253. Arquitecturas adicionales como LightGBM, CatBoost o redes neuronales temporales podrían mejorar aún más la detección de mora incipiente.
- **Temporalidad de los datos:** el *dataset* cubre obligaciones de los años 2024 y 2025, lo que limita la evaluación de la estabilidad del modelo ante cambios estructurales del entorno económico o climático. Se recomienda su reentrenamiento periódico a medida que se acumule nueva información.
- **Alcance geográfico:** los modelos fueron entrenados sobre la cartera de la Regional Occidente (Valle del Cauca, Cauca y Nariño). Su aplicación directa a otras regionales requiere validación con datos propios, dado que las dinámicas climáticas y de conflicto armado varían significativamente entre regiones del país.
- **Supuestos del modelo logístico ordinal:** el modelo polr asume proporcionalidad de las odds (*parallel regression assumption*), es decir, que el efecto de cada predictor es constante a través de los distintos umbrales de severidad. Aunque este supuesto fue evaluado durante la estimación, su posible violación parcial podría sugerir la exploración de modelos ordinales no proporcionales como alternativa futura.

6. Riesgo Crediticio Agropecuario (PowerBI)

El dashboard interactivo de riesgo crediticio agropecuario fue desarrollado en Microsoft Power BI Desktop como herramienta de visualización y monitoreo de los resultados del modelo predictivo. La herramienta integra en tres páginas diferenciadas los indicadores de cartera, los resultados de los modelos binomiales y los resultados del modelo ordinal multiclase, permitiendo la segmentación dinámica por oficina, género, destino de inversión, hecho victimizante, factor externo y ciudad de inversión mediante segmentadores interactivos.

6.1 Página Datos: Caracterización de la Cartera

La primera página del dashboard presenta una visión general de la cartera agropecuaria de la Regional Occidente. En la parte superior se muestran cuatro indicadores clave (KPI): el total de obligaciones (55.438), el saldo potencial (\$524 mil millones), el saldo total de cartera (\$387 mil millones) y el porcentaje de obligaciones vencidas (0,20%). Estos indicadores responden dinámicamente a los filtros aplicados por el usuario.

El componente geográfico muestra la distribución espacial de las obligaciones sobre un mapa interactivo del suroccidente colombiano, donde se evidencia la concentración de operaciones en los departamentos de Nariño y Cauca. El gráfico de barras horizontales “Créditos Departamento de Inversión” confirma esta distribución: Nariño lidera con 27.655 obligaciones, seguido de Cauca con 25.522 y Valle del Cauca con 2.106. En la franja inferior se presentan tres gráficos complementarios: la distribución por destino agrupado (donde Agricultura-cultivos concentra 35.690 obligaciones), las obligaciones por calificación interna (donde la calificación A agrupa 45.792 créditos, evidenciando una cartera predominantemente sana), y la distribución por estado de obligación mediante un gráfico de dona (88,07% vigente, 11,72% suspenso, 0,20% vencido).

La **Figura 26** presenta el dashboard interactivo de la página DATOS, donde se destacan los siguientes elementos complementarios: los segmentadores dinámicos ubicados en la parte superior permiten filtrar simultáneamente por Oficina, Género, Destino Agrupado, Hecho Victimizante, Factor Externo y Ciudad de Inversión, facilitando análisis territoriales y poblacionales específicos. La paleta de colores diferencia visualmente las categorías críticas: las barras en rojo intenso destacan los departamentos de mayor concentración (Nariño y Cauca), mientras que el gráfico de dona revela que la cartera en suspenso (11,72%) representa un segmento de atención prioritaria por su potencial de deterioro hacia mora efectiva. La calificación E, con 3.579 obligaciones, y la C, con 3.271, concentran el riesgo latente de impago, constituyendo en conjunto el 12,3% de la cartera con mayor probabilidad de migración hacia estados críticos.

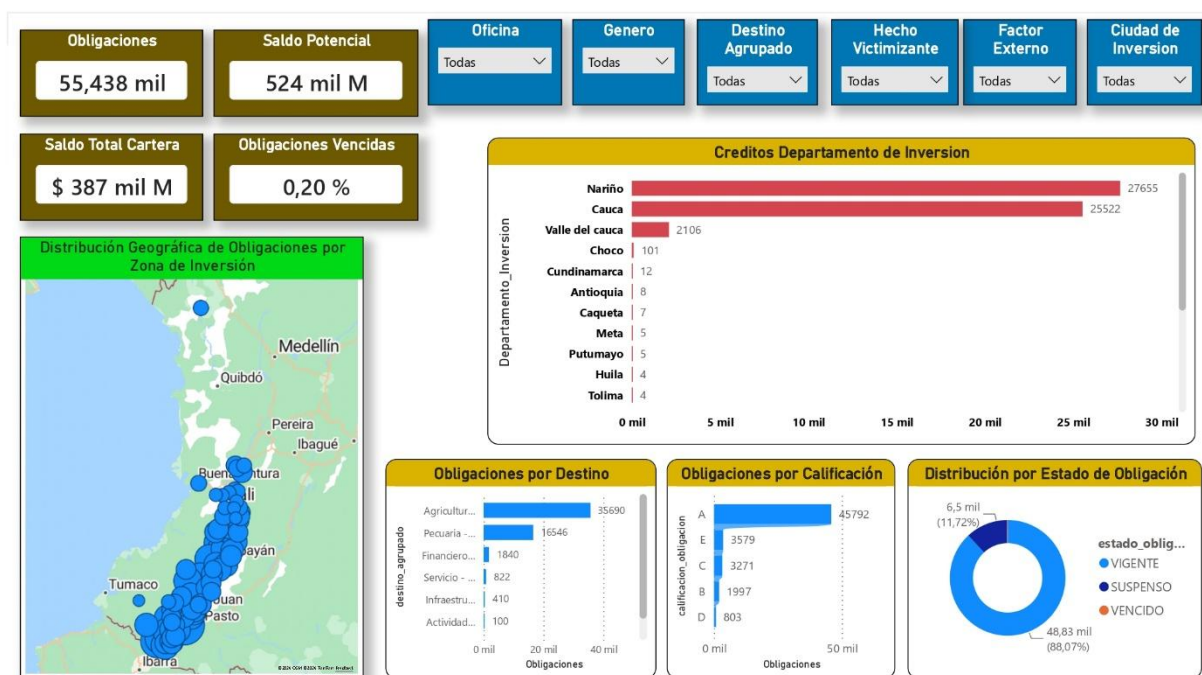


Figura 26: Dashboard interactivo — caracterización de la cartera agropecuaria Regional Occidente.

6.2 Página BINOMIAL: Resultados del Modelo Binario

La segunda página presenta los resultados de los seis modelos binomiales estimados (Regresión Logística, Árbol de Decisión y XGBoost, con y sin ponderación de clases). En la parte superior, un selector desplegable permite filtrar las métricas por modelo específico, mientras que cinco tarjetas resumen muestran las métricas globales: Especificidad (66,93%), Sensibilidad (11,59%), F1 (18,74%), Precisión (65,43%) y AUC (0,550). Dos botones de marcador permiten alternar entre las tablas de probabilidades de mora por factor externo climático y por hecho victimizante del conflicto armado, facilitando la consulta diferenciada por tipo de riesgo exógeno.

El gráfico de barras agrupadas “Métricas por modelo” permite comparar visualmente Precisión, Sensibilidad, Especificidad y Balanced Accuracy entre los seis modelos, evidenciando la superioridad del XGBoost ponderado en el equilibrio entre sensibilidad y especificidad. La curva ROC comparativa confirma este resultado, mostrando que el XGBoost ponderado y el Logit ponderado presentan las curvas más alejadas de la diagonal aleatoria. El gráfico de importancia de variables del XGBoost revela que el saldo, las cuotas pactadas y el monto desembolso son los predictores de mayor ganancia, seguidos por el tipo de factor externo y la tasa referencial, validando la relevancia de las variables exógenas climáticas y socioeconómicas en el modelo.

La **Figura 27** presenta el dashboard BINOMIAL, donde se observa que la Ola Invernal registra la

mayor probabilidad de mora en el modelo Logit ponderado (0,4685), seguida de Factores Naturales-Volcán (0,4659), mientras que el factor Económico presenta las probabilidades más bajas (0,0845), evidenciando que los fenómenos climáticos extremos generan un riesgo de mora significativamente superior al de las perturbaciones económicas. El promedio total de probabilidad de mora (0,4283 en Logit ponderado) sugiere que, bajo condiciones de estrés climático, aproximadamente 4 de cada 10 obligaciones presentan riesgo de incumplimiento.

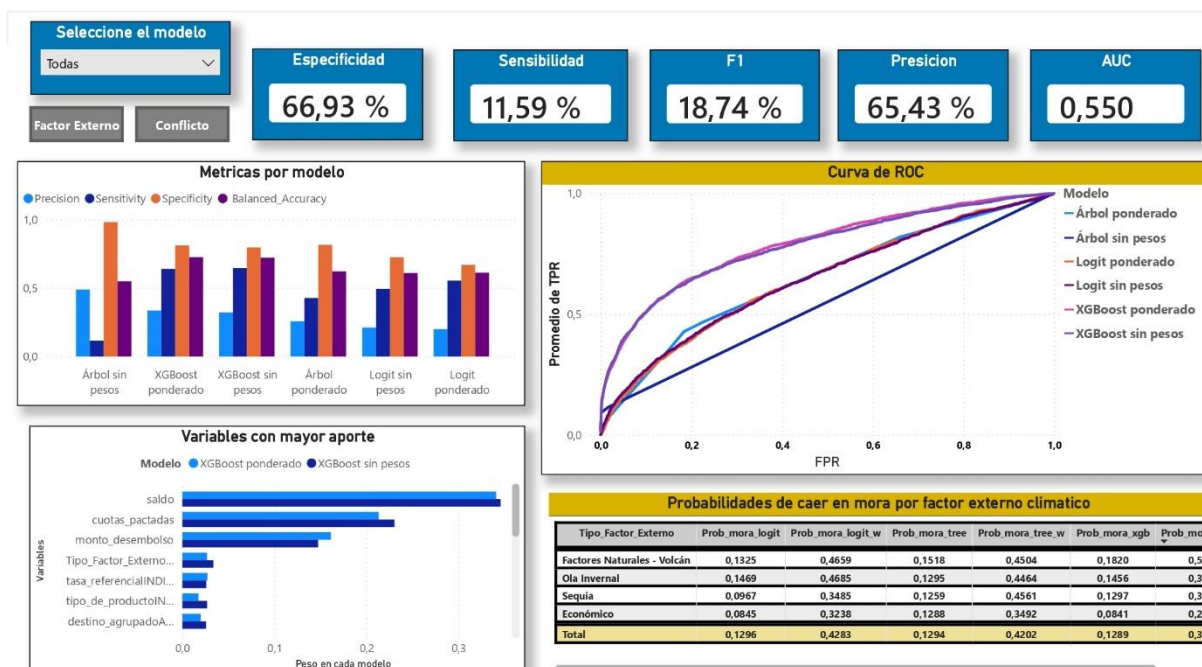


Figura 27: Dashboard interactivo — Página BINOMIAL: métricas comparativas de modelos binomiales, curva ROC e importancia de variables.

6.2 Resultados del Modelo de Severidad de Mora:

La tercera página presenta los resultados del modelo de regresión logística ordinal para la severidad de mora, clasificada en tres rangos temporales: 1 a 30 días, 31 a 90 días y mayor a 91 días. Doce tarjetas de métricas organizadas en tres filas muestran Especificidad, F1, Precisión y Sensibilidad para cada rango de mora, permitiendo identificar que el modelo presenta mayor capacidad discriminatoria en la mora temprana (Especificidad 1-30: 0,82) que en la mora crónica (Especificidad Mayor a 91: 0,62).

La tabla comparativa de modelos muestra las métricas macro (F1_macro, AUC_macro, BalAcc_macro) para los seis modelos evaluados. El Árbol ponderado lidera con AUC_macro de 13,15 y BalAcc_macro de 12,41, seguido del Árbol sin pesos. El gráfico de barras “Métricas por modelo” permite la comparación visual de las cuatro métricas entre modelos, evidenciando que la ponderación de clases mejora consistentemente la Sensibilidad a costa de una reducción moderada en la Especificidad. Este panel facilita

la selección informada del modelo óptimo según el criterio de gestión de riesgo priorizado por la entidad financiera.

La **Figura 28** presenta el dashboard MULTIMODAL, donde se observa que el F1 mejora progresivamente con la severidad de mora: 0,27 en mora temprana (1-30 días), 0,43 en mora intermedia (31-90 días) y 0,52 en mora crónica (mayor a 91 días), indicando que el modelo ordinal tiene mayor capacidad para identificar correctamente los casos de impago prolongado que los de mora incipiente. Adicionalmente, la Precisión sigue un patrón similar (0,31 $\rightarrow$ 0,44 $\rightarrow$ 0,57), sugiriendo que las predicciones de mora crónica son más confiables que las de mora temprana, lo cual tiene implicaciones directas para la priorización de acciones de cobranza preventiva en etapas iniciales del deterioro.



Figura 28. Página MULTIMODAL del dashboard: métricas de severidad de mora por rango temporal y comparación de modelos ordinales.

7. Conclusiones Y Trabajos Futuros

7.1 Conclusiones

El presente proyecto demostró que es posible construir modelos predictivos capaces de anticipar el deterioro de la cartera agropecuaria integrando variables crediticias, climáticas y de conflicto armado. A partir del análisis de 55.454 obligaciones de la Regional Occidente (Valle del Cauca,

Cauca y Nariño), se identificaron patrones significativos que permiten orientar la gestión proactiva del riesgo crediticio. A continuación, se presentan las principales conclusiones del estudio.

1. **Viabilidad del modelo predictivo integral.** El proyecto confirmó la viabilidad de integrar en un solo modelo predictivo variables crediticias internas (cuotas pactadas, destino del crédito, monto desembolso, saldo capital, modalidad de amortización) con variables exógenas de cambio climático (tipo de factor externo) y conflicto armado (hecho victimizante, *Comparativo_Ocurrencia_Hecho*). La inclusión de estas variables explícitamente en la estimación, con validación *out-of-sample* sobre el 30% del *dataset*, permitió obtener modelos con poder discriminatorio operacionalmente útil.
2. **Superioridad del XGBoost como motor de *scoring* binario.** El XGBoost ponderado se posicionó como el mejor modelo del portafolio con un AUC-ROC de 0.7921, Kappa de 0.3253 y F1 de 0.4400, superando al Logit Binomial (AUC = 0.6493, Kappa = 0.1462) y al Árbol de Decisión (AUC = 0.5779) en todas las métricas relevantes. Esta diferencia de aproximadamente 14 puntos en AUC respecto al mejor modelo lineal confirma la naturaleza no lineal e interactiva de los determinantes del riesgo crediticio agropecuario, que los modelos clásicos no logran capturar completamente. El XGBoost ponderado es el modelo recomendado para implementación operativa en sistemas de *scoring* masivo.
3. **El factor climático y el conflicto armado son predictores significativos del riesgo de mora.** El análisis de odds ratios del modelo Logit Binomial confirmó que la Ola Invernal aumenta el riesgo de mora en un 61% respecto a obligaciones clasificadas bajo factor económico (OR = 1.61), mientras que el desplazamiento y despojo eleva dicho riesgo en un 42% (OR = 1.42). El destino agrícola (agricultura y cultivos) incrementó el riesgo un 92% respecto a la ganadería (OR = 1.9). Estos hallazgos validan empíricamente la hipótesis central del proyecto: el cambio climático y el conflicto armado son determinantes estadísticamente significativos del deterioro de la cartera agropecuaria [18], y su inclusión en los modelos de riesgo mejora su relevancia operativa.
4. **Etapas tempranas del crédito como ventana crítica de riesgo.** El número de cuotas pactadas resultó uno de los predictores más significativos del modelo ($p < 0.001$): por cada desviación estándar adicional en las cuotas pactadas, la probabilidad de mora se reduce un 34.3% (OR = 0.657), indicando que los créditos de mayor plazo presentan menor riesgo relativo. Este hallazgo evidencia que los créditos de corto plazo y las

etapas iniciales del ciclo de vida constituyen la ventana de mayor exposición al riesgo, lo que justifica concentrar las estrategias de acompañamiento en ese período, especialmente para clientes con hecho victimizante o afectados por factores externos climáticos.

5. **Heterogeneidad territorial del riesgo.** El análisis reveló heterogeneidad significativa entre departamentos. Las obligaciones en Valle del Cauca tienen una probabilidad de mora un 34% mayor que las del Cauca ($OR = 1.341$, $p < 0.001$), mientras que las de Nariño presentan una probabilidad un 28% menor ($OR = 0.717$, $p < 0.001$). En el modelo ordinal, la Sequía se asocia con la mayor probabilidad de mora severa ($\beta = +1.732$, $p < 0.001$), mientras que los Factores Naturales-Volcán muestran el mayor efecto protector ($\beta = -2.345$, $p < 0.001$). Estos resultados evidencian la necesidad de políticas de colocación, seguimiento y gestión diferenciadas por territorio y tipo de factor externo.
6. **Valor de la ponderación de clases en *datasets* desbalanceados.** En todos los modelos estimados, la incorporación de pesos por clase produjo mejoras en la detección de mora (*Sensitivity*), con un *trade-off* controlado en Accuracy. Con un desbalance de 87.03% vs. 12.97% entre clases, la ponderación resultó crítica para evitar que los modelos sesgaran sus predicciones hacia la clase mayoritaria. El `XGBoost` ponderado logró el mejor equilibrio global (*Balanced Accuracy* = 0.7251), mientras que el Logit Ordinal Ponderado mejoró drásticamente la detección de mora temprana (*Sensitivity* 1–30 días: 0.034 $\rightarrow$ 0.614).
7. **Contribución a la gestión proactiva del riesgo.** Los modelos desarrollados permiten la construcción de sistemas de alerta temprana activados por eventos externos (olas invernales, orden público), la segmentación territorial diferenciada del riesgo y el diseño de estrategias por ciclo de vida del crédito. En conjunto, representan una base analítica sólida para fortalecer la toma de decisiones en la entidad financiera, con potencial de replicación en otras regionales del país previa validación con datos propios.
8. **Desarrollo de los ODS:** Este tipo de investigaciones marcan un precedente y es proyectar a la banca colombiana a desarrollar y aplicar los Objetivos de Desarrollo Sostenible planteados por las Naciones Unidas, la OCDE y la CEPAL [23] [24].

7.2 Trabajos Futuros

Los resultados obtenidos abren múltiples líneas de desarrollo futuro que permitirían ampliar el alcance, la precisión y la utilidad operativa del modelo predictivo:

Exploración de modelos de ensamble adicionales y redes neuronales. Se recomienda evaluar Random Forest, LightGBM, CatBoost y arquitecturas de redes neuronales tipo *LSTM* para capturar dinámicas temporales en el comportamiento de pago, dado que el *dataset* cubre periodos 2024–2025 con potencial de enriquecimiento longitudinal.

Incorporación de datos satelitales y geoespaciales de alta resolución. La integración de índices satelitales de vegetación (*NDVI*), datos de precipitación del IDEAM [7] a nivel municipal y cartografía de conflicto del CERAC con mayor desagregación temporal podría mejorar sustancialmente la capacidad predictiva del modelo y permitir un análisis de escenarios estacionales.

Desarrollo de un tablero interactivo de alertas tempranas. Se propone mejorar el *dashboard* en Power BI o Shiny que integre los modelos entrenados y permita actualizar automáticamente el perfil de riesgo de la cartera vigente ante la ocurrencia de eventos externos, facilitando la toma de decisiones en tiempo real por parte de los gestores de cartera regionales.

Reentrenamiento periódico y validación de estabilidad temporal. Dado que el *dataset* cubre únicamente los años 2024–2025, se recomienda establecer un protocolo de reentrenamiento semestral del modelo *XGBoost* con los datos más recientes, acompañado de pruebas de estabilidad de población (*PSI*, Population Stability Index) para detectar deriva conceptó y ajustar el umbral de clasificación según la evolución del entorno económico y climático.

Extensión a otras regionales y productos financieros. El marco metodológico desarrollado es replicable en otras regionales del país con las adaptaciones necesarias a sus dinámicas climáticas y de conflicto. Asimismo, se identifica como línea de trabajo futura la aplicación del modelo a la predicción de refinanciaciones, pérdidas dado el incumplimiento (*LGD*) y la valoración de instrumentos financieros derivados orientados a cobertura de riesgo agropecuario [11].

8. REFERENCIAS BIBLIOGRÁFICAS

- [1] S. M. Hsiang, M. Burke, and E. Miguel, "Quantifying the influence of climate on human conflict," *Science*, vol. 341, no. 6151, p. 1235367, 2013.
- [2] J. J. Perfetti, L. G. Caro, R. Junguito, and L. A. Zuleta, "Impacto de los factores de riesgo en el estado de la cartera del Banco Agrario," Fedesarrollo, 2015. [Online]. Available: <https://www.repository.fedesarrollo.org.co/handle/11445/2890>
- [3] J. Orrego Correa and S. Uribe Orozco, "Conflicto armado y sistema financiero: impacto del conflicto e iniciativas para la inclusión financiera en épocas de posconflicto," Universidad de los Andes, 2018.
- [4] J. Libreros Ruiz, "Conflicto armado en Colombia: evidencia de su impacto sobre la producción agrícola municipal," Universidad de los Andes, [2017].
- [5] C. Bohórquez-Peñuela, M. M. Gáfaró-González, K. Guzmán-Finol, and A. Pérez, "Climate risk and access to credit: Evidence from Colombian potato growers," *Borradores de Economía*, no. 1307, Banco de la República de Colombia, 2025. [Online]. Available: <https://doi.org/10.32468/be.1307>
- [6] IBM, "CRISP-DM help overview," IBM. [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>
- [7] IDEAM, "Informe nacional del estado del clima," Instituto de Hidrología, Meteorología y Estudios Ambientales, 2021.
- [8] DANE and UNGRD, "Reporte de afectaciones por eventos climáticos en el sector agropecuario," Departamento Administrativo Nacional de Estadística & Unidad Nacional para la Gestión del Riesgo de Desastres, 2022.
- [9] Centro Nacional de Memoria Histórica, "¡Basta ya! Colombia: memorias de guerra y dignidad," CNMH, 2020.
- [10] Asobancaria, "Integración de riesgos climáticos y factores ASG (Ambientales, Sociales y de Gobernanza) en el proceso de otorgamiento de créditos bancarios", Finanzas Sostenibles / Gestión de Riesgo Bancario, 2025.
- [11] FAO, "Modelos predictivos para la gestión del riesgo agropecuario en América Latina,"

Organización de las Naciones Unidas para la Alimentación y la Agricultura, 2020.

[12] La República, "Colombia, el tercer país más vulnerable al cambio climático," Feb. 24, 2014. [Online]. Available: <https://www.larepublica.co/archivo/colombia-el-tercer-pais-mas-vulnerable-al-cambio-climatico-2243441>

[13] Gobierno de Colombia, "Decreto 122 de 2022: Por el cual se modifica la estructura del Banco Agrario de Colombia S.A.," 2022. [Online]. Available: https://www.Bancoagrario.gov.co/system/files/2023-08/decreto_122_de_2022_0.pdf

[14] Banco Agrario de Colombia, "Código de ética y conducta (Código GA-CP-PL-001, Versión 2)," 2024. [Online]. Available: https://www.Bancoagrario.gov.co/system/files/2025-02/ga-cp-pl-001_vf.pdf

[15] Banco Mundial, "Colombia: Evaluación del Sistema Financiero," 2020. [Online]. Available: <https://www.Bancomundial.org>

[16] DNP - Departamento Nacional de Planeación, "Política Nacional de Gestión del Riesgo de Desastres y Adaptación al Cambio Climático," 2022.

[17] IDEAM, PNUD, and MADS, "Tercera Comunicación Nacional de Cambio Climático de Colombia," 2017.

[18] J. Rodríguez and C. Pineda, "Impacto del conflicto armado en la economía rural colombiana: Evidencia a nivel municipal," *Revista de Economía del Rosario*, vol. 24, no. 1, pp. 47-78, 2021.

[19] FAO, "Gestión del riesgo agroclimático en América Latina," 2019.

[20] Superintendencia Financiera de Colombia, "Informe de estabilidad del sistema financiero colombiano," 2023.

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed. Springer, 2021.

[22] R. J. T. Klein et al., "Climate change 2014: Impacts, adaptation, and vulnerability," IPCC, 2014.

[23] OCDE, "Colombia: Diagnóstico multidimensional," 2019.

- [24] CEPAL, "Territorios resilientes para una recuperación sostenible en América Latina y el Caribe," 2021.
- [25] IBM Corporation, *Guía de CRISP-DM de IBM SPSS Modeler*, versión 18, release 4, modificación 0, IBM Corp., 2018.
- [26] DANE (2023). Nota Estadística: Mujer rural y campesina en Colombia. Dirección de Metodología y Producción Estadística.
- [27] Circular Externa 022 de 2008 de la SFC, que incorpora el Anexo III “Modelo de Referencia de Cartera Comercial – MRC”, mediante el cual se regulan las categorías de riesgo crediticio, la probabilidad de incumplimiento (PI) y la pérdida dado el incumplimiento (PDI)
- [28] Escobar Andrés J. "Bancarización en Peligro: Conflicto Armado e Inclusión Financiera en Colombi," Universidad de los Andes, [2017].
- [29] Superintendencia Financiera de Colombia (SFC) y el Grupo Banco Mundial (World Bank Group), " *Un realismo no tan mágico: Pruebas de esfuerzo climático de la banca colombiana*", 2021.
- [30] Rojas-Jimenez, K. 2022. Ciencia de Datos para Ciencias Naturales. [Online]. Available: https://bookdown.org/keilor_rojas/CienciaDatos/
- [31] James, G., Witten, D., Hastie, T., y Tibshirani, R., "An Introduction to Statistical Learning: with Applications in R", 2.^a ed., Springer, 2021.
- [32] James, G., Witten, D., Hastie, T., Tibshirani, R., y Taylor, J., "An Introduction to Statistical Learning: with Applications in Python", Springer, 2023.
- [33] Unidad Nacional de Protección (UNP) y Departamento Administrativo Nacional de Estadística (DANE), "Diagnóstico de la Política Pública para la atención y prevención de personas dadas por desaparecidas: Informe Suroccidente Colombiano", Bogotá: SNB, 2025.
- [34] Congreso de la República de Colombia, "Ley 1448 de 2011: Por la cual se dictan medidas de atención, asistencia y reparación integral a las víctimas del conflicto armado interno y se dictan otras disposiciones", Diario Oficial No. 48.096, 2011.

[35] Instituto de Estudios para el Desarrollo y la Paz (Indepaz), "Violencia en Colombia. Informe Anual 2023," Observatorio de DDHH y Conflictividades, Indepaz, Bogotá, dic. 2023. [Online]. Available: <https://indepaz.org.co/violencia-en-colombia-informe-anual-2023>

[36] El Espectador, "Cauca: los frentes del EMC y el ELN que operan en el departamento," El Espectador, 2024. [Online]. Available: <https://www.elespectador.com>

[37] Banco de la República de Colombia, "Informe de Estabilidad Financiera 2024-II," Bogotá: Banco de la República, 2024. [Online]. Available: <https://www.banrep.gov.co>

Anexos

1. Diccionario de los datos

Nombre de la variable	Descripción	Tipo de dato	Ejemplo
Fecha corte	Fecha en la que se procesa la información de la cartera de la entidad	Fecha	31/07/2025
regional	Nombre que se le otorga al agrupamiento de varios departamentos	Texto	Cauca, Nariño y Valle del Cauca
zonal	Nombre que se le otorga al agrupamiento de varios municipios (oficinas) por departamento	Texto	para este caso son 7 zonales, Cauca centro y norte , cauca sur. Nariño centro, nariño norte y nariño sur. Zona pacífica

nombre_oficina	Nombre de la oficina donde se originó el crédito	Texto	para este caso la regional occidente cuenta con 95 oficinas, donde las zonales con mas oficinas son pacifica y valle
Longitud	Sistema de coordenadas que describen la ubicación de un un punto en la tierra	Numérico	-76,85607
Latitud			1,913429
id	Identificación del cliente (NIT o Cedula) para el caso de colombia	Numérico	100001
nombre_cliente	Nombre del cliente (puede estar anonimiz.)	Texto	—
numero_Banco	Número único de la operación de crédito	Numérico	4610152690620
dias_vencido	Número de días de mora de la obligación	Numérico	0 días, entre 1 - 90 días
rango	Clasificación de los días vencidos	Texto	0 días, 1a 90 días, mayor a 90 días y mayor a 120 días
fecha_prox_ven	Fecha del próximo vencimiento de la cuota	Fecha	10/08/2025
saldo	Saldo capital de la obligación (millones de pesos)	Numérico	1,973,683

Banca_cliente	Segmento de banca del cliente	Texto	Microfinanzas, Agropecuaria, Personal Pyme consolidada, Empresarial, Oficial
tipo_cartera	especificacion del tipo de producto que tiene un cliente	Texto	Cartera, tarjeta y sobregiros
genero	Género del cliente	Texto	Femenino, Masculino
ciudad_inversion	Ciudad donde se ejecuta la inversión	Texto	Toribio, Chachagui, Tuquerres
fecha_inicio_op	Fecha de inicio de la operación del credito	Fecha	11/01/2023
monto_desembolso	Monto total desembolsado	Numérico	10,000,000.00
calificacion_obligacion	Calificación de riesgo de la obligación	Texto (A–E)	A, B, C, D y E
modalidad	Modalidad del crédito	Texto	MV, SV, AV, TV,DV,DV
valor_vencido	Valor vencido acumulado	Numérico	0
saldo_cuota_prox_venc	Valor de la próxima cuota	Numérico	424,070
etapa	Etapas de seguimiento del crédito de acuerdo a los días de mora o políticas del Banco	Texto	Preventivo, administrativo, prejudicio y juridico
estado_obligacion	Estado de la obligación, su descripción depende los días de mora y estado de cobranza	Texto	Vigente sin deuda, Vigente, Suspenso y Vencido

cuotas_pagadas	Número de cuotas ya pagadas	Numérico	30
cuotas_pactadas	Número total de cuotas pactadas al momento de la originacion	Numérico	35
Cero_cuotas_pagadas	Indicador si no ha pagado ninguna cuota	Texto (S/N)	Si o No
mo_reestructurada	Si el crédito ha sido reestructurado	Texto (S/N)	Si o No
num_reestructuraciones	Número de veces que se ha reestructurado	Numérico	1
hecho_victimizante	Tipo de hecho victimizante (conflicto armado)	Texto	Pérdida de Bienes, etc.
fecha_ven_op	Fecha de vencimiento de la obligación	Fecha	10/12/2025
destino_agrupado	Agrupación general del destino del crédito	Texto	Microfinanzas, Café
tipo_de_producto	Producto solicitado por el cliente	Texto	Capital trabajo para café
tipo_de_productor	Tipo de producto financiero	Texto	Microfinanzas, Pequeños productores, Medianos productores, Empresarial cadena agriopecuaria, Grandes productores finagro y No agropecuario

tipo_identificacion	Tipo de documento de identificación	Texto	CE, CC
tasa_referencial	Tasa referencial usada en la obligación	Texto	TASA Cero, Indicador Bancario
F_OCURRENCIA_HECHO	Fecha de ocurrencia del hecho victimizante	Fecha	27/04/2010
COMPARATIVO_OCURRENCIA_HECHO	Comparar si el hecho ocurrió antes de otorgar el crédito o después	Texto	V-D Víctimas Antes del desembolso V-D Víctima después del desembolso
Desplazado	Indica si el cliente fue desplazado	Texto (S/N)	Si o No
Factor Externo	Factor externo que afecta la operación	Texto	Ola Invernal
Tipo de Factor Externo	Clasificación del factor externo	Texto	CAFÉ, CLIMÁTICO, etc.
Vigencia_Factor	Vigencia del factor externo	Fecha	31/07/2024

2. Código Python — Análisis Exploratorio de Datos y Preprocesamiento

Archivo: *tesis.html* — Notebook Jupyter exportado a HTML

Herramienta: Python 3 · Google Colaboratory

Autores: Flavio Alejandro Castillo Perea · Johan Sebastián Valencia Cárdenas

Acceso local: <file:///C:/Users/johan/Downloads/tesis.html>

Descripción del contenido:

Este anexo contiene el notebook completo de Python desarrollado en Google Colaboratory, que incluye: (i) carga y depuración del *dataset* de cartera agropecuaria; (ii) ajuste de nombres de variables y recodificación de categorías; (iii) análisis exploratorio univariado y bivariado con visualizaciones; (iv) y (v) exportación del *dataset* limpio (*df_limpio.xlsx*) utilizado como insumo de los modelos R. Librerías principales: pandas, numpy, matplotlib, seaborn, scikit-learn, funpymodeling.

Instrucciones de acceso:

Abra el archivo tesis.html en cualquier navegador web (Chrome, Edge, Firefox). No requiere instalación adicional.

3. Código R — Análisis de Correspondencias Múltiples (MCA)

Archivo: *RMD_MCA.html* — R Markdown exportado a HTML

Herramienta: R · RStudio · Paquetes: *FactoMineR*, *factoextra*, *tidyverse*

Autores: Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea

Publicado en RPubs: <https://rpubs.com/josevaca/1433757>

Descripción del contenido:

Este anexo contiene el código R Markdown completo del Análisis de Correspondencias Múltiples (MCA) aplicado sobre las variables categóricas del *dataset* de cartera agropecuaria. El MCA permitió identificar las asociaciones más relevantes entre variables y orientó la selección de predictores para los modelos de riesgo crediticio. Se incluyen los biplots de dimensiones, la inercia explicada y el ranking de contribución de cada variable al modelo.

Instrucciones de acceso:

Abra el archivo RMD_MCA.html en cualquier navegador web. No requiere instalación adicional.

4. Código R — Modelos Logit Binomial, Árbol de Decisión y XGBoost

Archivo: *LOGIT_BINOMIAL_CON_SIN_PESO--I-RMD.html*

Herramienta: R · RStudio · Paquetes: *caret*, *pROC*, *rpart*, *xgboost*, *tidyverse*

Autores: Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea

Publicado en RPubs: <https://rpubs.com/josevaca/1430493>

Descripción del contenido:

Este anexo contiene el código R Markdown completo de los modelos de clasificación binaria (Al Día vs. En Mora) sobre el *dataset* de cartera agropecuaria ($n = 55.454$ obligaciones). Incluye: (i) Regresión Logística Binomial sin y con ponderación de clases (*glm*, familia binomial, umbral Youden); (ii) Árbol de Decisión sin y con pesos (*rpart*, $cp = 0.01$); (iii) XGBoost sin y con pesos (*xgb.train*, $\eta = 0.05$, $max_depth = 4$, 800 rondas, *early stopping* 20 rondas). Para cada modelo se reportan matrices de confusión, AUC-ROC, Accuracy, Kappa, *Sensitivity*, *Specificity*, F1 y *Balanced Accuracy* sobre el conjunto de prueba (*test* 30%, $n = 16.635$). Semilla de reproducibilidad: `set.seed(2930)`.

Instrucciones de acceso:

Disponible en línea en RPubs (<https://rpubs.com/josevaca/1430493>) o abriendo el archivo HTML adjunto en cualquier navegador web.

5. Código R — Modelo Logístico Ordinal (Severidad de Mora)

Archivo: *SEVERIDAD_MORA.html* — R Markdown exportado a HTML

Herramienta: R · RStudio · Paquetes: MASS, caret, pROC, rpart, xgboost, tidyverse

Autores: Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea

Publicado en RPubs: <https://rpubs.com/josevaca/1431363>

Descripción del contenido:

Este anexo contiene el código R Markdown completo de los modelos de severidad de mora implementados sobre las obligaciones en mora ($n = 7.191$; $train = 5.035$, $test = 2.156$). La variable objetivo es ordinal con tres categorías: 1–30 días, 31–90 días y 91+ días. Se implementaron seis modelos: Logit Ordinal (polr), Árbol de Decisión y XGBoost multiclase, cada uno con y sin ponderación de clases (proporción objetivo 30/35/35). Se reportan métricas por clase (*Sensitivity*, *Balanced Accuracy*) y AUC macro. Semilla: `set.seed(2930)`.

Instrucciones de acceso:

Disponible en línea en RPubs (<https://rpubs.com/josevaca/1431363>) o abriendo el archivo HTML adjunto en cualquier navegador web.

6. Tabla Comparativa Interactiva de Modelos

Archivo: *tabla_comparativa_modelos_v2.html* — Tabla HTML interactiva

Herramienta: HTML5 · CSS3 · JavaScript

Acceso Sharepoint: tabla_comparativa_modelos_v2.html

Autores: Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea

Descripción del contenido:

Este anexo presenta la tabla comparativa interactiva de métricas de desempeño para los ocho modelos evaluados, organizada en cuatro secciones: Logit Binomial, Logit Ordinal (polr), Árbol de Decisión (rpart) y XGBoost, cada uno con y sin ponderación de clases. La tabla permite identificar visualmente el mejor y peor valor por columna mediante marcadores de color. Todos los datos corresponden a los resultados sobre el conjunto de prueba ($test$ 30%) sin *data leakage*.

Instrucciones de acceso:

Abra el archivo *tabla_comparativa_modelos_v2.html* en cualquier navegador web. No requiere instalación adicional.

7. Dashboard Interactivo de Riesgo Crediticio Agropecuario - PowerBI

Archivo: *Dashboard_Riesgo_Agropecuario.pbix*

Acceso Sharepoint: [DASHBOARD - MAESTRIA CIENCIA DE DATOS.pbix](#)

Herramienta: Microsoft Power BI Desktop

Autores: Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea

Descripción del contenido:

Este anexo corresponde al *dashboard* interactivo desarrollado en Microsoft Power BI para la visualización y monitoreo del riesgo crediticio agropecuario en la Regional Occidente (Valle del Cauca, Cauca y Nariño). El *dashboard* integra indicadores clave de cartera (mora, saldos, departamentos), alertas basadas en factores externos climáticos y de conflicto armado, y los resultados de los modelos predictivos. Permite la segmentación territorial y por tipo de productor, facilitando la toma de decisiones proactiva por parte de los gestores de cartera regionales.

Instrucciones de acceso:

Para la versión interactiva, abra el archivo .pbix en Microsoft Power BI Desktop (versión 2024 o superior). Las capturas de las principales vistas se adjuntan en formato PDF como referencia.

8. Reporte Analítico de Riesgo Crediticio Agropecuario — Regional Occidente

Archivo: *Reporte Analítico Riesgo Agropecuario.pdf*

Acceso Sharepoint: [Reporte Analítico Riesgo Agropecuario.pdf](#)

Herramienta: *Adobe Acrobat / cualquier visor de PDF*

Autores: *Johan Sebastián Valencia Cárdenas · Flavio Alejandro Castillo Perea*

Descripción del contenido:

Este anexo corresponde al reporte analítico de riesgo crediticio agropecuario desarrollado como complemento del dashboard interactivo. El reporte sintetiza los hallazgos clave del modelo predictivo para la Regional Occidente (Valle del Cauca, Cauca y Nariño), incluyendo indicadores clave de cartera, perfil de riesgo por departamento, impacto de factores climáticos y de conflicto armado, odds ratios del modelo seleccionado, matriz de alertas tempranas y recomendaciones estratégicas para la gestión proactiva de la cartera agropecuaria.

Instrucciones de acceso:

Abra el archivo PDF en cualquier visor (Adobe Acrobat, navegador web). El documento está disponible en SharePoint en el enlace indicado arriba.