Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Software

Desarrollo de un modelo de Machine Learning usando el algoritmo XGBoost para medir la probabilidad de impago de clientes en una entidad bancaria, 2026

TESIS

Para optar el Título Profesional de Ingeniero de Software

AUTOR

Eduardo Rafael JÁUREGUI ROMERO

ASESOR

Dr. Santiago Domingo Moquillaza Henríquez

Lima, Perú

2026



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Jáuregui, E. (2026). *Desarrollo de un modelo de Machine Learning usando el algoritmo XGBoost para medir la probabilidad de impago de clientes en una entidad bancaria, 2026*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Software]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Eduardo Rafael Jáuregui Romero
Tipo de documento de identidad	DNI
Número de documento de identidad	73954094
URL de ORCID	https://orcid.org/0009-0008-0248-2601
Datos de asesor	
Nombres y apellidos	Santiago Domingo Moquillaza Henríquez
Tipo de documento de identidad	DNI
Número de documento de identidad	08280889
URL de ORCID	https://orcid.org/0000-0001-9531-881X
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Lenis Rossi Wong Portillo
Tipo de documento	DNI
Número de documento de identidad	10438282
Miembro del jurado 1	
Nombres y apellidos	Javier Antonio Prudencio Vidal
Tipo de documento	DNI
Número de documento de identidad	09858276
Datos de investigación	
Línea de investigación	C.0.3.3. Desarrollo de modelos y aplicación de las tecnologías de información y comunicaciones.

Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Edificio: Facultad de Ingeniería de Sistemas e Informática (FISI) País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Calle: Ciudad Universitaria (Av. Carlos Germán Amézaga N° 375 / Av. Venezuela cdra. 34) Latitud: -12.05928 Longitud: -77.08051
Año o rango de años en que se realizó la investigación	Marzo 2024 - Julio 2026
URL de disciplinas OCDE	**Ciencias de la computación** https://purl.org/pe-repo/ocde/ford#1.02.01 **Estadísticas, Probabilidad** https://purl.org/pe-repo/ocde/ford#1.01.03 **Negocios, Administración** https://purl.org/pe-repo/ocde/ford#5.02.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Software

Acta de Sustentación de Tesis

Siendo las once (11) horas del día veinte ocho (28) del mes octubre de año 2026, se reunieron en la sala presidido por la Dra. Lenis Rossi Wong Portillo, el Mg. Javier Antonio Prudencio Vidal (Miembro) y el Dr. Santiago Domingo Moquillaza Henríquez (Asesor), para la sustentación de la tesis Intitulada: ***"DESARROLLO DE UN MODELO DE MACHINE LEARNING USANDO EL ALGORITMO XGBOOST PARA MEDIR LA PROBABILIDAD DE IMPAGO DE CLIENTES EN UNA ENTIDAD BANCARIA, 2026"****, por el Bachiller* ***Eduardo Rafael, JÁUREGUI ROMERO****, para obtener el Título Profesional de Ingeniero de Software.*

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a responder las preguntas formuladas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez las preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de dieciocho (18).

A continuación, el Presidente del Jurado Dra. Lenis Rossi Wong Portillo, declara al Bachiller Ingeniero de Software.

Siendo 12 horas, se levantó la sesión.

Dra. Lenis Rossi Wong Portillo
Presidente

Mg. Javier Antonio Prudencio Vidal
Miembro

Dr. Santiago Domingo Moquillaza Henríquez
Asesor

UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

Anexo 1
CERTIFICADO DE SIMILITUD

Yo, SANTIAGO DOMINGO MOQUILLAZA HENRIQUEZ en mi condición de asesor acreditado con el Dictamen de la EPS N° 000004-2026-EPISW-FISI/UNMSM de la tesis, cuyo título es: **"DESARROLLO DE UN MODELO DE MACHINE LEARNING USANDO EL ALGORITMO XGBOOST PARA MEDIR LA PROBABILIDAD DE IMPAGO DE CLIENTES EN UNA ENTIDAD BANCARIA, 2026"** presentado por el bachiller: JÁUREGUI ROMERO, Eduardo Rafael, para optar el título profesional de Ingeniero de Software. CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 6 % de similitud, nivel PERMITIDO para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título profesional de Ingeniero de Software.

Lima, 17 de noviembre 2026

Firma del Asesor:

Nombres y Apellidos del asesor: Dr. SANTIAGO DOMINGO MOQUILLAZA HENRIQUEZ
DNI: 08280889

* Según corresponda título profesional

La vigencia del presente certificado corresponde a un período de noventa (90) días desde la fecha de su emisión; siempre que no haya observaciones en el texto durante el proceso de revisión por el jurado informante o surja alguna contingencia.

DEDICATORIA

A mi familia por su apoyo incondicional, compañía y ejemplo de perseverancia. A mis padres, fuente constante de amor y motivación en cada momento de esfuerzo y reflexión

Eduardo Rafael Jauregui Romero

AGRADECIMIENTOS

Expreso mi más sincero agradecimiento al Dr. Santiago Domingo Moquillaza Henríquez por su valiosa guía, paciencia y dedicación a lo largo de todo este proceso. Asimismo, agradezco a los docentes de la Facultad de Ingeniería de Sistemas e Informática por sus aportes y sugerencias que enriquecieron significativamente este trabajo. Mi gratitud también a la Universidad Nacional Mayor de San Marcos, por la formación recibida que hizo posible este proyecto.

Por último, a mi familia y seres queridos, por su apoyo constante, comprensión y motivación en cada etapa.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SOFTWARE

DESARROLLO DE UN MODELO DE MACHINE LEARNING USANDO EL ALGORITMO XGBOOST PARA MEDIR LA PROBABILIDAD DE IMPAGO DE CLIENTES EN UNA ENTIDAD BANCARIA, 2026

Autor: EDUARDO RAFAEL JAUREGUI ROMERO

Asesor: MOQUILLAZA HENRIQUEZ SANTIAGO DOMINGO

Titulo: Tesis para optar el Título Profesional de Ingeniero de Software

Fecha: Junio del 2026

RESUMEN

La morosidad en préstamos personales de la banca peruana alcanzó niveles críticos durante la pandemia del COVID-19. En julio de 2020, la deuda acumulada de los morosos en este producto llegó a 9,235 millones de soles. Esta investigación desarrolló un modelo de aprendizaje automático basado en XGBoost para medir la probabilidad de impago en una entidad bancaria peruana, incorporando interpretabilidad mediante SHAP, monitoreo continuo y un agente basado en LLM, alineado con la normativa SBS.

La metodología fue cuantitativa, con diseño no experimental, sobre 1,735,340 clientes con préstamos vigentes entre junio de 2021 y octubre de 2022. Un conjunto de 582 variables fue reducido a 48 mediante filtros estadísticos, análisis bivariado y transformación WOE. Siete algoritmos fueron evaluados con validación cruzada estratificada. XGBoost resultó el mejor. Tras optimización bayesiana de hiperparámetros en nube, el modelo alcanzó un Gini de 68.29

%, un AUC de 84.15 % y un KS de 0.51 en el conjunto fuera de tiempo, todos clasificados como buenos según escalas del sector bancario.

El análisis SHAP identificó el porcentaje de uso de tarjeta de crédito en el último mes, el mínimo de pasivos totales en los últimos tres meses y el máximo de días de atraso en el sistema financiero como las variables de mayor peso explicativo. El monitoreo registró un Gini de casi 70 puntos en abril de 2025 y un PSI de 7.1 %, sin alertas de deriva. El agente integrado, basado en RAG, respondió consultas en lenguaje natural sobre el comportamiento del modelo, facilitando la interpretación de métricas para equipos técnicos y de negocio.

Palabras clave: riesgo crediticio, aprendizaje automático, probabilidad de impago, préstamos personales, interpretabilidad de modelos, modelos de lenguaje grande

NATIONAL MAJOR UNIVERSITY OF SAN MARCOS

FACULTY OF SYSTEMS ENGINEERING AND INFORMATICS

PROFESSIONAL SCHOOL OF SOFTWARE ENGINEERING

DEVELOPMENT OF A MACHINE LEARNING MODEL USING THE XGBOOST ALGORITHM TO MEASURE THE PROBABILITY OF CLIENT DEFAULT IN A BANKING ENTITY, 2026

Author: EDUARDO RAFAEL JAUREGUI ROMERO

Advisor: MOQUILLAZA HENRIQUEZ SANTIAGO DOMINGO

Title: Tesis para optar el Título Profesional de Ingeniero de Software

Date: June 2026

ABSTRACT

Non-performing loans in the personal credit portfolio of Peruvian banks reached critical levels during the COVID-19 pandemic. In July 2020, the accumulated debt of delinquent borrowers in this product amounted to 9,235 million soles. This study developed a machine learning model based on XGBoost to measure the probability of default in a Peruvian banking institution, incorporating SHAP-based interpretability, continuous monitoring, and an LLM-based agent, aligned with SBS regulations.

A quantitative methodology with a non-experimental design was applied to 1,735,340 customers with active loans between June 2021 and October 2022. An initial dataset of 582 variables was reduced to 48 through statistical filters, bivariate analysis, and WOE

transformation. Seven algorithms were evaluated using stratified cross-validation. XGBoost emerged as the best candidate. After Bayesian hyperparameter optimization on cloud infrastructure, the model achieved a Gini coefficient of 68.29 %, an AUC of 84.15 %, and a KS statistic of 0.51 on the out-of-time dataset, all classified as good according to banking industry benchmarks.

SHAP analysis identified credit card utilization in the last month, minimum total liabilities in the last three months, and maximum days past due in the financial system as the variables with the highest explanatory weight. The monitoring system recorded a Gini coefficient of nearly 70 points in April 2025 and a PSI of 7.1 %, with no data drift alerts. The integrated RAG-based agent answered natural language queries about model behavior, facilitating metric interpretation for technical and business teams.

Keywords: credit risk, machine learning, probability of default, personal loans, model interpretability, large language models

ÍNDICE DE CONTENIDO

AGRADECIMIENTOS III
RESUMEN IV
ABSTRACT VI
ÍNDICE DE CONTENIDO VIII
ÍNDICE DE FIGURAS XII
ÍNDICE DE TABLAS XV
ÍNDICE DE ECUACIONES XVII
INTRODUCCIÓN 1
CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO 3
1.1 ANTECEDENTES DEL PROBLEMA 3
1.1.1 Antecedentes Internacionales 3
1.1.2 Antecedentes Nacionales 7
1.2 PLANTEAMIENTO DEL PROBLEMA 10
1.3 FORMULACIÓN DEL PROBLEMA 13
1.3.1 Problema General 13
1.3.2 Problema Específicos 13
1.4 OBJETIVOS DE LA INVESTIGACIÓN 14
1.4.1 Objetivo General 14
1.4.2 Objetivos Específicos 14
1.5 JUSTIFICACIÓN DE LA INVESTIGACIÓN 14
1.5.1 Justificación Teórica 15
1.5.2 Justificación Práctica 15
1.6 ALCANCE 16
1.7 LIMITACIONES DE LA INVESTIGACIÓN 17
1.7.1 Espacial 17
1.7.2 Temporal 18
1.7.3 Social 18
1.8 IDENTIFICACIÓN DE VARIABLES 18
1.8.1 Variable Independiente (X) 18
1.8.2 Variables predictoras 19
1.8.3 Variable Dependiente (y) 19
1.9 UNIVERSO Y MUESTRA 19
1.9.1 Universo 19
1.9.2 Muestra 20
1.10 METODOLOGÍA Y DISEÑO DE LA INVESTIGACIÓN 21
1.10.1 Tipo de Investigación 21
1.10.2 Nivel de Investigación 22
1.10.3 Enfoque de Investigación 23
1.10.4 Diseño de Investigación 23
1.10.5 Complemento metodológico: Ciencias del diseño (DSR) 24
1.11 ORGANIZACIÓN DE LA TESIS 25

CAPÍTULO II: MARCO TEÓRICO 27
2.1 RIESGO CREDITICIO 27
2.1.1 Préstamos Personales 27
2.1.2 Morosidad y Default 27
2.1.3 Scoring Crediticio 28
2.2 MACHINE LEARNING 28
2.2.1 Aprendizaje Supervisado 29
2.2.2 Clasificación Binaria y Balanceo de Clases 29
2.2.3 Algoritmos de Clasificación 30
2.2.3.1 Regresión Logística 30
2.2.3.2. Árboles de Decisión 31
2.2.3.3 Random Forest 31
2.2.3.4 AdaBoost 32
2.2.3.5 CatBoost 32
2.2.3.6 LightGBM 33
2.2.3.7 XGBOOST 34
2.2.4 Métricas de Evaluación 35
2.2.4.1 Curva ROC y AUC 35
2.2.4.2 Coeficiente de Gini 36
2.2.4.3 Estadístico de Kolmogorov–Smirnov (KS) 37
2.2.4.4 Matriz de Confusión 37
2.2.4.5 Sensibilidad y Especificidad 38
2.2.4.6. Índice de Youden 38
2.2.5 Interpretabilidad de Modelos 39
2.2.5.1 Importancia de la Interpretabilidad 39
2.2.5.2 SHAP (Shapley Additive exPlanations) 40
2.2.5.3. Peso de Evidencia (WOE) 41
2.2.5.4 Análisis Bivariado 42
2.2.6 Monitoreo y Gestión de Riesgo de Modelos 43
2.2.6.1 Monitoreo de Modelos 43
2.2.6.2 CRISP-ML(Q) 43
2.2.6.3 Índice de Estabilidad Poblacional (PSI) 44
2.2.6.4 Monitoreo de Performance 45
2.2.6.5 Gestión de Riesgo de modelos y marco normativo SBS 46
CAPÍTULO III: ESTADO DEL ARTE METODOLÓGICO 47
3.1 INTRODUCCIÓN 47
3.2 BUSQUEDA DE PAPERS 47
3.3 PROCESO DE RECOLECCIÓN 49
3.4 RESULTADOS DE LA RECOLECCIÓN 50
3.5 DESARROLLO DEL ESTADO DEL ARTE 53
3.6 HALLAZGOS CLAVES 59
CAPÍTULO IV: IMPLEMENTACIÓN 65
4.1 DESARROLLO DE LA SOLUCIÓN 65

4.1.1 Construcción del dataset inicial con variables de carácter crediticio, social y comportamental 66
4.1.1.1 Recolección de la Poblacion Objetivo y Definición del Horizonte Temporal. 66
4.1.1.2 Construcción y Definición de la Variable Objeto (Target) mediante Analisis de Roll Rates 68
4.1.1.3 Aplicación de Filtros de Inclusión y Exclusión a la Población Objetivo 71
4.1.1.4 Evaluación de la Estabilidad de la Tasa de Malos 73
4.1.1.5 Recolección y Distribución de Variables por Fuentes de Información 74
4.1.2 Preprocesamiento, limpieza y análisis exploratorio del dataset inicial 75
4.1.2.1 Filtros Generales de Reducción de Dimensionalidad (Missings, Gini Mínimo y Correlación) 75
4.1.2.2 Análisis Bivariado y Transformación de Variables mediante Peso de la Evidencia (WOE) 77
4.1.2.3 Filtro de Importancia Relativa mediante el Modelo de Prueba (Toy Model) .. 87
4.1.2.4 Tratamiento de Valores Nulos 88
4.1.2.5 Detección y Aislamiento de Valores Atípicos (Outliers) mediante Isolation Forest 90
4.1.2.6 Análisis de Correlacion del Dataset Final 94
4.1.3 Evaluación, entrenamiento y selección del mejor modelo predictivo de impago para Préstamos Personales 95
4.1.3.1 Estrategia de Partición y Sampleo Estratificado del Dataset 95
4.1.3.2 Entrenamiento y Validación Cruzada de Algoritmos Candidatos 96
4.1.3.2.1 Regresión Logística 97
4.1.3.2.2 Árbol de Decisión 100
4.1.3.2.3 Random Forest 102
4.1.3.2.4 AdaBoost Classifier 104
4.1.3.2.5 CatBoost Classifier 107
4.1.3.2.6 LightGBM Classifier 109
4.1.3.2.7 XGBOOST Classifier 112
4.1.3.3 Comparativa de Modelos Candidatos (Benchmarking) 114
4.1.3.4 Optimización Bayesiana de Hiperparámetros mediante Arquitectura Cloud en Amazon SageMaker 115
4.1.3.5 Evaluación de Performance del Modelo Final (Métricas ROC, Gini, KS y Matriz de Confusión) 118
4.1.3.6 Transformación a Score Crediticio (Scorecard) 123
4.1.4 Diseño e implementación del sistema de seguimiento de performance, estabilidad e interpretabilidad del modelo 124
4.1.4.1 Análisis de Interpretabilidad Global e Individual mediante Valores de Shapley (SHAP) 124
4.1.4.2 Arquitectura en Cloud para el Cálculo de Indicadores de Monitoreo 131
4.1.4.3 Despliegue de Plataforma Web de Monitoreo: Model Risk Management Suite 134
4.1.4.3.1 Flujo de Operaciones 135
4.1.4.3.2 Arquitectura Interna del Frontend (SPA) 136

4.1.4.3.3 Arquitectura del Backend Serverless 137
4.1.4.4 Agente Model Risk Management (MRM) basado en Modelos de Lenguaje (LLM) y LangChain 137
4.1.4.4.1 Arquitectura del Agente: Integración del Patrón ReAct y LangChain 138
4.1.4.4.2 Tools del Agente: Despliegue de herramientas y acceso a datos del proyecto 138
4.1.4.4.3 Mecanismo de Recuperación de Información y Generación Aumentada (RAG) 140
4.1.4.4.4 Despliegue del Agente como Microservicio 142
4.2. VALIDACIÓN DEL SISTEMA 144
4.2.1 Validación del Artefacto del Modelo de Probabilidad de Impago 145
4.2.2 Validación de la Plataforma WEB – Model Risk Management Suite 146
4.2.2.1 Monitoreo del Modelo Performance y Estabilidad 146
CAPÍTULO V: DISCUSIÓN, CONCLUSIONES Y RECOMENDACIONES 154
5.1 DISCUSIÓN 154
5.2. CONCLUSIONES 157
5.2.1. Conclusión general 158
5.2.2. Conclusiones específicas 158
5.3. RECOMENDACIONES 160
5.3.1. Recomendaciones para la entidad bancaria 160
5.3.2 Recomendaciones para futuras investigaciones 161
REFERENCIAS 163
ANEXOS 181

ÍNDICE DE FIGURAS

FIGURA 1. EVOLUCIÓN DE DEUDA Y EL NÚMERO DE MOROSOS EN AÑO 2019-2020 EN EL PERÚ 11
FIGURA 2. DISTRIBUCIÓN DE CANTIDAD DE MOROSOS Y DEUDA FINANCIERA POR PRODUCTO CREDITICIO 12
FIGURA 3. DIAGRAMA GENERAL DEL UNIVERSO-UNIDAD DE OBSERVACIÓN DE CLIENTES DE LA ENTIEDAD BANCARIA..... 19
FIGURA 4. FLUJO PARA LA OBTENCIÓN DE LA MUESTRA..... 21
FIGURA 5. CONSTRUCCION DE ARBOL DEL ALGORITMO CATBOOST..... 33
FIGURA 6. CONSTRUCCION DE ARBOL DEL ALGORITMO LIGHTGBM..... 34
FIGURA 7. CONSTRUCCION DE ARBOL DEL ALGORITMO XGBOOST..... 35
FIGURA 8. EJEMPLO DE DESCOMPOSICIÓN SHAP PARA LA INTERPRETABILIDAD..... 39
FIGURA 9. APORTE LOCAL DE VARIABLES PARA PREDICCIÓN FINAL EN SHAP 41
FIGURA 10. CALCULO DE WOE PARA UNA VARIABLE CONTINUA..... 42
FIGURA 11. FLUJO DE METODOLOGIA CRISP-ML(Q)..... 44
FIGURA 12. CALCULO DE PSI Y CATEGORIZACION DE DATA DRIFT 45
FIGURA 13. PROCESO DE RECOLECCION DE PAPERS FINALISTAS 49
FIGURA 14. LAS 6 FASES DE METODOLOGIA CRISP-ML(Q) 65
FIGURA 15. ESTRUCTURA PARA RECOLECCIÓN DE FUENTES..... 67
FIGURA 16. NUMERO DE CLIENTES CON PRESTAMOS PRESONAES DURANTE 2021-2022 67
FIGURA 17. DIAGRAMA DE CONSTRUCCIÓN Y COMPORTAMIENTO DEL TARGET 68
FIGURA 18. CURVA DE TASA DE DETERIORO – ROL RATES 70
FIGURA 19. EVOLUTIVO DE TASA DE MALOS POR MES 73
FIGURA 20. FUENTES CONSULTADAS PARA LA CREACIÓN DEL DATASET 74
FIGURA 21. FLUJO DE FITROS PARA DATASET INICIAL DE VARIABLES..... 75
FIGURA 22. GRAFICO BIVARIADO PARA LA VARIABLE PROCEDENCIA PROFESIONAL 86
FIGURA 23. GRAFICO BIVARIADO PARA LA VARIABLE SEGMENTACIÓ RFM ..87
FIGURA 24. PRINT DE OBJETO OPTIMAL BINNING DE PROCEDENCIA..... 87
FIGURA 25. GRAFICO DE VARIABLES MISSING CALCULADAS CON LIBRERÍA MISSINGNO 89
FIGURA 26. MAPA DE CALOR CORRELACIONAL PARA LOS MISSING DE LAS VARIABLES. 89
FIGURA 27. VISUALIZACIÓN DE ANOMALIAS POR INSTANCIA..... 91
FIGURA 28. DETECCIÓN DE ANOMALÍAS EN VARIABLES TOP 92
FIGURA 29. DISTRIBUCIÓN DE TARGET POR REGISTROS NORMALES Y ANÓMALOS 92
FIGURA 30. GRAFICO DISPERSIÓN DE OUTLIERS USANDO PCA..... 93

FIGURA 31. MAPA DE CALOR DE CORRELACION DE SET DE VARIABLES FINALISTAS 94
FIGURA 32. FLUJO DE ENTRENAMIENTO GENERAL PARA BENCHMARKING DE MODELOS 96
FIGURA 33. COEFICIENTES DEL ALGORITMO REGRESIÓN LOGÍSTICA 98
FIGURA 34. CURVA ROC PARA EL ALGORITMO REGRESIÓN LOGÍSTICA 99
FIGURA 35. CURVA ROC PARA EL ALGORITMO ÁRBOL DE DECISIÓN 101
FIGURA 36. TOP VARIABLES MAS IMPORTANES DEL ÁRBOL DE DECISIÓN ...101
FIGURA 37. CURVA ROC PARA EL ALGORITMO RANDOM FOREST 103
FIGURA 38. TOP VARIABLES PARA EL ALGORITMO RANDOM FOREST 104
FIGURA 39. CURVA ROC PARA EL ALGORITMO ADABOOST CLASIFIER.....106
FIGURA 40. TOP VARIABLES PARA EL ALGORITMO ADABOOST CLASIFIER ..106
FIGURA 41. CURVA ROC PARA EL ALGORITMO CATBOOST CLASIFIER 108
FIGURA 42. TOP VARIABLES PARA EL ALGORITMO CATBOOST CLASIFIER...109
FIGURA 43. CURVA ROC PARA EL ALGORITMO LIGHTGBM CLASSIFIER 110
FIGURA 44. PESO DE VARIABLES PARA EL ALGORITMO LIGHTGBM CLASIFIER 111
FIGURA 45. CURVA ROC PARA EL ALGORITMO XGBOOST CLASSIFIER 113
FIGURA 46. TOP VARIABLES MAS IMPORTANTES PARA EL ALGORITMO XGBOOST CLASSIFIER 113
FIGURA 47. DIAGRAMA DE ARQUITECTURA PARA EL FLUJO DE EJECUCION Y OPTIMIZACION DEL MODELO.....116
FIGURA 48. CURVA ROC DEL MODELO XGBOOST OPTIMIZADO.....119
FIGURA 49. SELECCIÓN DE PUNTO DE CORTE OPTIMO – MÉTODO DE YOUDEN 121
FIGURA 50. MATRIZ DE CONFUSIÓN – VISTA POR CUADRANTES 122
FIGURA 51. FORCE PLOT PARA UN REGISTRO DEL CONJUNTO OOT 127
FIGURA 52. GRÁFICO DE DEPENDENCIAL PARCIAL VARIABLE TOP 1- % USO TC ÚLTIMO MES.....128
FIGURA 53. GRÁFICO DE DEPENDENCIAL PARCIAL VARIABLE TOP 2 – MINIMO PASIVOS U3M 129
FIGURA 54. GRÁFICO DE DEPENDENCIAL PARCIAL VARIABLE TOP 3 – MAXIMO DIAS DE ATRASO RIESGO DIRECTO PERSONA NATURAL U3M 129
FIGURA 55. GRÁFICO DE RESUMEN DE TENDENCIA SHAP PARA TOP 20 VARIABLES 130
FIGURA 56. DIAGRAMA DE ARQUITECTURA SERVELESS PARA LA ORQUESTACIÓN Y CÁLCULO DE INDICADORES DE MONITOREO.....132
FIGURA 57. TABLA DE RESULTADOS DE PERFORMANCE CONSULTADA MEDIANTE ATHENA PARA INDICADOR GINI.....134
FIGURA 58. DIAGRAMA GENERAL DE LA PLATAFORMA WEB 134
FIGURA 59. DIAGRAMA DE ARQUITECTURA SEPARADO POR CAPAS. 136
FIGURA 60. EJEMPLO DE AGENTE BASADO EN PATRÓN REACT.....138
FIGURA 61. DIAGRAMA DE ALTO NIVEL DE FUNCION DE HERRAMIENTAS DE RECUPERACIÓN.....140

FIGURA 62. PROCESO DE INGESTA DE DATOS A BASE DE DATOS VECTORIAL 141
FIGURA 63. PROCESO DE RECUPERACIÓN DE DATOS PARA RAG.....142
FIGURA 64. DESPLIEGUE DEL AGENTE EN AMAZON WEB SERVICES (AWS) ..142
FIGURA 65. FLUJO DE INTERACCIÓN COMPLETO DEL AGENTE Y ESTRUCTURA GENERAL 143
FIGURA 66. CONFORME DE LA VALIDACIÓN DEL ARTEFACTO DEL MODELO DE PROBABILIDAD DE IMPAGO.....146
FIGURA 67. VISTA DE PERFORMANCE GENERAL DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE.....147
FIGURA 68. VISTA DE TM POR DECIL DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE 147
FIGURA 69. VISTA DE INDICADORES DE MC DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE.....148
FIGURA 70. COMPARATIVO DE TASA DE MALOS PARA VARIABLE TOP 1 DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE.....149
FIGURA 71. VISTA ESTABILIDAD GENERAL DEL MODELO DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE.....150
FIGURA 72. VISTA ESTABILIDAD GENERAL DEL MODELO DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT SUITE.....151
FIGURA 73. PSI POR VARIABLES DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT.....151
FIGURA 74. DISTRIBUCIÓN DE LA VARIABLE CON ALERTA AMBAR DISPONIBLE EN LA PLATAFORMA MODEL RISK MANAGEMENT 152

ÍNDICE DE TABLAS

TABLA 1. *QUERIES Y FILTROS USADOS EN LA BUSQUEDA DEL ESTADO DEL ARTE* 48
TABLA 2. *CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN DE LOS PAPERS* 48
TABLA 3. *FUENTES POR TRABAJOS SELECCIONADOS* 50
TABLA 4. *LISTA DE PAPERS SELECCIONADOS* 51
TABLA 5. *RESUMEN DE HALLAZGOS CLAVES EN PAPERS* 62
TABLA 6. *VACIOS IDENTIFICADOS EN PAPERS.* 64
TABLA 7. *FASES DE METODOLOGIA CRISP-ML(Q) RELACIONADAS A LOS OBJETIVOS ESPECIFICOS* 66
TABLA 8. *ANALISIS DE ROLL RATES – MATRIZ DE TRANSICIÓN ENTRE INICIO Y FIN* 69
TABLA 9. *FILTROS APLICADOS A LA POBLACION OBJETIVO – CRITERIOS DE NEGOCIO Y CALIDAD*.....71
TABLA 10. *FILTRO CASCADA DE LA POBLACION OBJETIVO* 72
TABLA 11. *PESO DE VARIABLES POR FUENTE DE INFORMACIÓN*.....75
TABLA 12. *SET DE VARIABLES DISGREGADO POR GRUPO Y FILTRO APLICADO* 76
TABLA 13. *DICCIONARIO DE VARIABLES FILTRADAS Y GRAFICO BIVARIADO* 77
TABLA 14. *TRANSFORMACIÓN WOE PARA LA VARIABLE PROCEDENCIA PROFESIONAL* 85
TABLA 15. *TRANSFORMACIÓN WOE PARA LA VARIABLE SEGMENTACIÓN RFM* .86
TABLA 16. *HIPERPARAMETROS DEL TOY MODEL* 88
TABLA 17. *TOP VARIABLES CON MAYOR Y MENOR IMPORTANCIA DEL TOY MODEL* 88
TABLA 18. *PARAMETROS DEL INSOLATION FOREST PARA AL BUSQUEDA DE OUTLIERS* 91
TABLA 19. *PARAMETROS DEL OBJETO TRAIN_TEST_SPLIT*.....96
TABLA 20. *PARAMETROS DEL STRATIFIEDKFOLD PARA LA REGRESIÓN LOGÍSTICA* 97
TABLA 21. *RESULTADOS DEL KFOLD PARA LA REGRESIÓN LOGÍSTICA* 98
TABLA 22. *PARAMETROS DEL KFOLD PARA ALGORITMO ÁRBOL DE DECISIÓN* 100
TABLA 23. *PARAMETROS PARA ALGORITMO RANDOM FOREST*.....102
TABLA 24. *PARAMETROS PARA ALGORITMO ADABOOST CLASSIFIER* 105
TABLA 25. *RESULTADOS KFOLD PARA ALGORITMO ADABOOST CLASSIFIER* ...105
TABLA 26. *PARAMETROS PARA ALGORITMO CATBOOST CLASSIFIER*.....107
TABLA 27. *RESULTADOS KFOLD PARA ALGORITMO CATBOOST CLASSIFIER* ...108
TABLA 28. *PARAMETROS PARA ALGORITMO LIGHTGBM CLASSIFIER* 109
TABLA 29. *RESULTADOS PARA ALGORITMO LIGHTGBM CLASSIFIER*.....110
TABLA 30. *PARAMETROS PARA ALGORITMO XGBOOST CLASSIFIER*.....112
TABLA 31. *RESULTADOS PARA ALGORITMO XGBOOST CLASSIFIER*.....113
TABLA 32. *RESUMEN DE RESULTADOS OBTENIDOS POR CADA MODELO.* 114

TABLA 33. *DETALLE DE PERIODOS USADOS EN ENTRENAMIENTO Y OOT* 116
TABLA 34. *ESPACIO DE BÚSQUEDA Y VALOR ENCONTRADO PARA LA OPTIMIZACIÓN DE HIPERPARAMETROS*.....117
TABLA 35. *RESUMEN DE INDICADORES DE PERFORMANCE DEL MODELO XGBOOST OPTIMIZADO*.....119
TABLA 36. *CLASIFICACIÓN DEL GINI SEGÚN ESCALAS*.....120
TABLA 37. *CLASIFICACIÓN DEL KS SEGÚN ESCALAS* 121
TABLA 38. *METRICAS DERIVADAS DE LA MATRIZ DE CONFUSIÓN*.....122
TABLA 39. *TOP VARIABLES MAS IMPORTANTES DE MODELO XGBOOST FINAL* 124
TABLA 40. *TOOLS DEL AGENTE MODEL RISK MANAGEMENT* 139
TABLA 41. *TABLA LIKERT* 153
TABLA 42. *CRITERIOS EVALUADOS* 153

ÍNDICE DE ECUACIONES

ECUACIÓN 1: CALCULO DEL GINI 36
ECUACIÓN 2: REGRESIÓN LOGISTICA 97
ECUACIÓN 3: SCALE POS WEIGHT 117
ECUACIÓN 4: ESTADÍSTICO DE KOLMOGOROV-SMIRNOV (KS) 120
ECUACIÓN 5: TRANSFORMACIÓN A SCORE CREDITICIO (SCORECARD) 123

INTRODUCCIÓN

La gestión del riesgo crediticio es una de las funciones más críticas de la banca minorista. En préstamos personales, esa criticidad se acentúa: los plazos son largos, las garantías son escasas y el seguimiento individual de la cartera es difícil de sostener con métodos tradicionales (Toledo Concha, 2022). Cuando ese seguimiento falla, las consecuencias son directas: morosidad creciente, rentabilidad deteriorada y menos crédito disponible para nuevos clientes.

La pandemia del COVID-19 hizo visible esa vulnerabilidad. En el Perú, la deuda acumulada de los morosos en préstamos personales llegó a S/9,235 millones en julio de 2020. la morosidad de los préstamos personales aumentó 3 puntos porcentuales durante 2020. Al cierre de ese año, los morosos de préstamos personales y tarjetas de crédito representaron, en conjunto, el 37 % del total de morosos (Gestión, 2021). Este no es un problema exclusivamente peruano. En toda Latinoamérica, los préstamos personales figuran entre los productos más demandados y, por tanto, entre los que más exponen a las instituciones financieras (Infobae, 2022).

La literatura reciente muestra que los algoritmos de machine learning en particular los métodos de gradient boosting como XGBoost, LightGBM y CatBoost superan de forma consistente a los modelos estadísticos tradicionales en la predicción del impago crediticio (Brygała y Korol, 2024; Aruleba y Sun, 2024). Sin embargo, su adopción en entornos bancarios regulados enfrenta dos obstáculos: la interpretabilidad de las decisiones y el mantenimiento del rendimiento en producción.

En el contexto peruano, Hermitaño Castro (2022) identificó que los bancos locales reconocen la superioridad predictiva de estos algoritmos, pero señalan el problema de la "caja negra" como barrera para su implementación bajo los requisitos de la SBS. La Resolución

S.B.S. N° 00053-2023 formaliza esa exigencia al establecer criterios de gobierno corporativo, interpretabilidad y monitoreo continuo. Hasta la fecha, no existe en la literatura nacional un trabajo que integre estos tres componentes en un modelo de impago para la cartera de préstamos personales.

Esta investigación cubre ese vacío mediante un modelo XGBoost para medir la probabilidad de impago en una entidad bancaria peruana, con horizonte de 12 meses. El modelo se desarrolla aplicando CRISP-ML(Q).

La propuesta incorpora tres elementos diferenciadores: (1) preprocesamiento riguroso con WOE, Isolation Forest y reducción de 582 a 48 variables; (2) interpretabilidad global e individual mediante SHAP; y (3) monitoreo continuo en nube con indicadores Gini, KS y PSI, accesibles vía plataforma web y agente basado en LLM.

La investigación es aplicada, con enfoque cuantitativo y diseño no experimental, sobre 1,735,340 clientes con préstamos vigentes entre 2021 y 2022.

El documento se organiza en cinco capítulos. El Capítulo I presenta el planteamiento metodológico. El Capítulo II desarrolla el marco teórico. El Capítulo III sistematiza el estado del arte. El Capítulo IV detalla la implementación y validación del modelo. El Capítulo V expone la discusión, conclusiones y recomendaciones.

CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO

1.1 Antecedentes del Problema

Para situar bien esta investigación, conviene revisar los antecedentes tanto a nivel internacional como nacional en el campo de la predicción de riesgo crediticio con modelos de machine learning. Este recorrido permite identificar las tendencias más relevantes, los algoritmos que se han probado con éxito y los enfoques metodológicos que otras instituciones financieras y centros de investigación han adoptado. De esta forma se construye un marco de referencia sólido que justifica y orienta el desarrollo del modelo predictivo que se propone en este trabajo.

1.1.1 Antecedentes Internacionales

En los últimos años se ha publicado una cantidad considerable de estudios sobre la aplicación de machine learning a la predicción del riesgo crediticio. A continuación se revisan los trabajos más relevantes que abordan problemas como el desbalance de clases, la interpretabilidad, la equidad y el uso de datos alternativos.

Chen et al. (2024), en su artículo “Interpretable machine learning for imbalanced credit scoring datasets” publicado en el *European Journal of Operational Research*, estudiaron cómo afecta el desbalance de clases a la estabilidad de los métodos de explicabilidad LIME y SHAP. Observaron que cuando la proporción de eventos raros cae por debajo del 5 %, la estabilidad de estas explicaciones disminuye de manera notable. Este resultado es directamente aplicable al contexto bancario, donde las tasas de morosidad suelen ser bajas y es necesario manejar con cuidado el desbalance para que las interpretaciones del modelo resulten confiables.

Hurlin et al. (2026), en “The Fairness of Credit Scoring Models” (*Management Science*), se centraron en el problema de la equidad (*fairness*) en los modelos de scoring. Demostraron que los sesgos pueden provenir tanto de los datos de entrenamiento como del

propio algoritmo, y propusieron un marco formal para detectarlos y corregirlos. Su trabajo subraya que no basta con evaluar la precisión predictiva; también es necesario considerar el impacto social y ético de las decisiones automatizadas en el sector financiero.

Djeundje et al. (2021), en "Enhancing credit scoring with alternative data" (*Expert Systems with Applications*), exploraron el valor de incorporar variables no tradicionales, como información de correo electrónico y mediciones psicométricas. Sus experimentos mostraron que estas variables mejoran la capacidad predictiva, especialmente en personas que no tienen historial crediticio formal. Este hallazgo resulta particularmente útil para el caso peruano, donde una parte importante de la población aún no está bancarizada.

De Lange et al. (2022), mencionan en "Explainable AI for credit assessment in banks" (*Journal of Risk and Financial Management*), compararon un modelo LightGBM explicado con SHAP frente a la regresión logística tradicional que utilizaba un banco. El enfoque basado en LightGBM + SHAP no solo obtuvo mejor rendimiento predictivo, sino que también facilitó el cumplimiento de los requisitos regulatorios de explicabilidad. El estudio demuestra que es posible usar modelos más complejos en entornos bancarios sin perder la transparencia que exigen los supervisores.

Moldovan (2023), en "Algorithmic Decision Making Methods for Fair Credit Scoring" (*IEEE Access*), realizó una comparación amplia de doce técnicas de mitigación de sesgos. Evaluó cada método con métricas de equidad, exactitud balanceada y rentabilidad, y ofreció un marco práctico para tomar decisiones algorítmicas más justas en el otorgamiento de crédito.

Alonso Robisco & Carbó Martínez (2022), en "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction" (*Financial Innovation*), propusieron evaluar los modelos no solo por su precisión, sino también por su riesgo de modelo. Midieron tanto la estabilidad estadística de las predicciones como el tiempo que se necesita para explicar las decisiones con SHAP. Su trabajo pone de relieve que la

interpretabilidad y la estabilidad son componentes tan importantes como la precisión en el contexto bancario.

Liu et al. (2024), en "Credit scoring prediction leveraging interpretable ensemble learning" (*Journal of Forecasting*), confirmaron que los modelos de ensamble suelen superar a los clasificadores individuales en tareas de scoring crediticio. Entre los algoritmos probados, CatBoost mostró el mejor desempeño general. Además, utilizaron SHAP para ofrecer interpretaciones tanto globales (variables más importantes) como locales (explicación de predicciones individuales), lo que permite mantener alto poder predictivo sin sacrificar la transparencia exigida en entornos regulados.

Aruleba & Sun (2024), en "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation" (*IEEE Access*), compararon varios clasificadores de ensamble en conjuntos de datos clásicos de crédito. XGBoost destacó en el conjunto German Credit, mientras que Random Forest lo hizo en el Australian. Los autores combinaron técnicas de remuestreo como SMOTE-ENN con explicaciones SHAP y demostraron que es posible alcanzar alta precisión manteniendo la interpretabilidad del modelo.

Hlongwane et al. (2024), en "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data" (*PLOS ONE*), incorporaron variables como información de redes sociales, calificación económica regional y datos de población local. Esta inclusión elevó el AUC hasta 0.79360, lo que refuerza la tendencia actual de aprovechar fuentes de datos no tradicionales para mejorar la predicción en poblaciones con poco o ningún historial crediticio.

Brygała & Korol (2024), en "Personal bankruptcy prediction using machine learning techniques" (*Economics and Business Review*), compararon varios algoritmos para predecir bancarrota personal. LightGBM, CatBoost y XGBoost fueron los que ofrecieron mejor rendimiento, consolidando la evidencia sobre la ventaja de los métodos de boosting en problemas de clasificación crediticia.

Han et al. (2024), al estudiar “NOTE: Non-parametric oversampling technique for explainable credit scoring” (*Scientific Reports*), propusieron una técnica de sobremuestreo no paramétrica (NOTE) como alternativa al SMOTE tradicional. Señalaron que SMOTE puede introducir solapamiento y ruido al generar muestras sintéticas, mientras que su método mejora la exactitud y la estabilidad. Este trabajo muestra que el desarrollo de nuevas técnicas de remuestreo sigue siendo un área activa de investigación.

Han & Jung (2024), en su articulo “NATE: Non-pArameTric approach for explainable credit scoring on imbalanced class” (*PLOS ONE*), evaluaron distintas estrategias de remuestreo en datos desbalanceados. Las técnicas de sobremuestreo, especialmente combinadas con algoritmos de gradient boosting, superaron al submuestreo y mejoraron métricas como AUC, MCC y F1. Sus resultados refuerzan la necesidad de tratar adecuadamente el desbalance de clases en contextos crediticios.

Tigges et al. (2024), en su investigación “Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data” (*Technological Forecasting and Social Change*), realizaron un análisis cualitativo con expertos del sector fintech. Los entrevistados reconocieron mejoras en la predicción, reducción de la morosidad e inclusión de poblaciones no bancarizadas, pero también advirtieron sobre riesgos como sesgos algorítmicos, falta de consentimiento informado y problemas de trazabilidad en las decisiones automatizadas.

En conjunto, la literatura internacional revisada muestra una evolución clara: se ha pasado de modelos tradicionales a enfoques basados en ensambles y técnicas de boosting como XGBoost, LightGBM y CatBoost. Al mismo tiempo, ha crecido la preocupación por la interpretabilidad (principalmente mediante SHAP), el manejo correcto del desbalance de clases y el aprovechamiento de datos alternativos. Este panorama sirve de base para la presente investigación, que propone desarrollar un modelo XGBoost para estimar la probabilidad de

impago en una entidad bancaria peruana, incorporando interpretabilidad y mecanismos de monitoreo continuo alineados con la normativa de la SBS.

1.1.2 Antecedentes Nacionales

En el Perú, el interés por el machine learning aplicado al riesgo crediticio ha crecido rápido. Los trabajos locales siguen las mismas líneas que los internacionales, pero con un añadido: la normativa de la SBS exige transparencia y explicabilidad.

Hermitaño Castro (2022), en su estudio "Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática", evaluó la literatura para determinar ventajas y desventajas de los algoritmos de inteligencia artificial en el riesgo crediticio. Sus resultados señalaron que, si bien estos algoritmos ofrecen mayor precisión, herramientas como las redes neuronales sufren del problema de "caja negra", lo que dificulta su explicabilidad ante los entes regulatorios. Por ello, recomendó enfáticamente el uso de modelos híbridos o la integración de métodos de interpretabilidad como SHAP para garantizar la transparencia y el cumplimiento normativo en el sector financiero peruano.

Gonzalez Manrique & Coveñas De León (2024), en su investigación "Construcción de un score crediticio mediante el análisis de datos alternativos y métodos de inteligencia artificial para impulsar la inclusión financiera", desarrollaron un modelo de evaluación crediticia enfocado en la población no bancarizada, empleando variables no convencionales. Al comparar diversos modelos, sus resultados mostraron que XGBoost fue el más robusto y efectivo, con una precisión de 0.96 y un área bajo la curva ROC de 0.95. Este hallazgo demostró que la integración de datos alternativos e inteligencia artificial mejora significativamente la exactitud en la medición del riesgo crediticio, especialmente en segmentos poblacionales con historial financiero limitado.

Aradiel Castañeda et al. (2025), en su estudio "Explainable Machine Learning for Credit Card Default Prediction Using Web-Scraped Financial Data: A Case Study in the

Peruvian Banking Sector", desarrollaron un modelo predictivo para anticipar el riesgo de impago en tarjetas de crédito utilizando información recopilada mediante web scraping de los principales bancos peruanos. Sus hallazgos evidenciaron que LightGBM superó a otros enfoques, con una exactitud del 89.4% y un AUC de 0.94. Asimismo, confirmaron que la aplicación de técnicas como SHAP permite explicar las decisiones del modelo, facilitando su adopción como herramienta estratégica en entornos financieros regulados y contribuyendo a la democratización del crédito en el país.

Suxo (2024), en su tesis "Implementación de una solución de machine learning para la asignación de créditos financieros en una entidad financiera", abordó el despliegue de un modelo computacional diseñado para apoyar la toma de decisiones sobre la viabilidad de créditos otorgados a clientes top o preferenciales. Sus resultados demostraron que el uso de machine learning para estimar ingresos potenciales logró reducir notablemente las demoras en las evaluaciones tradicionales e incrementó la flexibilidad en las tasas de interés y montos ofrecidos, adaptándose mejor a escenarios volátiles y ambiguos propios del mercado peruano.

Perales (2024), en su trabajo "Modelo de calificación basado en Redes Neuronales para la predicción de Riesgo Crediticio: Caja Financiera", propuso un modelo para discernir entre clientes con buen y mal perfil de pago en el sector de las microfinanzas. Los resultados de su investigación evidenciaron que la arquitectura de red neuronal multicapa implementada, con ajuste adaptativo de pesos, alcanzó una precisión predictiva del 91.5%, superando ampliamente otros enfoques estadísticos y confirmando la alta eficacia de las redes neuronales para mitigar la morosidad en el contexto de las cajas financieras peruanas.

Vasquez (2025), en su investigación "Análisis comparativo de algoritmos de machine learning para predecir morosidad en clientes afiliados a la entidad financiera San Francisco de Mocupe", analizó distintas técnicas para optimizar la gestión de crédito ante la falta de información precisa. Su evaluación de desempeño demostró que LightGBM obtuvo la mayor

precisión (95%), seguido por XGBoost (94%) y Random Forest (93%), posicionando a LightGBM como la herramienta más eficiente para identificar a clientes morosos y proteger la liquidez institucional en el ámbito de las microfinanzas regionales.

López (2019), en su estudio "Comparación de modelos de aprendizaje de máquina en la predicción del incumplimiento de pago en el sector de las microfinanzas", evaluó diversos algoritmos de clasificación aplicando técnicas de balanceo de datos como SMOTE y codificación One Hot Encoding. Sus resultados indicaron que XGBoost presentó el mejor desempeño en condiciones desbalanceadas, con una exactitud de 97.53% y un F1-Score superior al de redes neuronales y regresión logística, lo que ratifica la fortaleza de los métodos de boosting en contextos financieros reales con alta desproporción entre clases.

Zegarra Valdivia & Gil Aguilar (2025), en su publicación "Elaboración de perfiles y el scoring crediticio: caracterización jurídica y aplicación del régimen legal peruano de protección de datos personales", analizaron los retos legales de la evaluación algorítmica en Perú. Su análisis concluyó que el scoring crediticio automatizado produce efectos jurídicos directos en los usuarios, por lo que activa de inmediato las obligaciones de "tratamiento objetivo" según el Reglamento de la Ley de Protección de Datos Personales, exigiendo a las entidades justificar la fuente de sus datos, delimitar roles y garantizar principios de transparencia y proporcionalidad.

Zavala & Poma (2024), en su trabajo "Implementación de un Dashboard inteligente para el análisis predictivo del riesgo crediticio en portafolios de inversión de un holding financiero", propusieron una solución que centraliza y evalúa la información utilizando machine learning. Al testear distintos modelos, sus resultados evidenciaron que XGBoost fue el algoritmo de mayor utilidad, con una precisión del 90%, un recall del 85% y un AUC-ROC de 0.92, lo que aseguró la confiabilidad de las proyecciones y optimizó la capacidad analítica de los gestores de inversión.

En el terreno normativo, la Superintendencia de Banca, Seguros y AFP (2023), mediante la Resolución S.B.S. N° 00053-2023, aprobó el Reglamento de Gestión de Riesgos de Modelo. Esta norma establece los criterios formales, los lineamientos de gobierno corporativo y los requerimientos mínimos de interpretabilidad que las entidades deben aplicar en el desarrollo, validación y uso de cualquier modelo algorítmico o de inteligencia artificial enfocado al riesgo crediticio. Constituye el marco regulatorio de referencia para esta investigación.

En conjunto, estos antecedentes nacionales confirman la creciente adopción de técnicas de machine learning en el sector financiero peruano, con una clara superioridad de los algoritmos de boosting como XGBoost y LightGBM en la predicción del riesgo crediticio. También se observa un énfasis creciente en la interpretabilidad de los modelos mediante herramientas como SHAP, en el manejo adecuado del desbalance de clases y en la incorporación de datos alternativos para mejorar la precisión predictiva. La normativa SBS, representada por la Resolución N° 00053-2023, exige transparencia, validación y monitoreo continuo de los modelos utilizados por las entidades financieras. Esta revisión proporciona un marco de referencia nacional que orienta y fundamenta la presente investigación, en la cual se propone desarrollar un modelo XGBOOST para medir la probabilidad de impago de clientes en una entidad bancaria peruana, incorporando elementos de interpretabilidad y monitoreo continúo alineados con la normativa vigente.

1.2 Planteamiento del problema

Los préstamos personales son uno de los productos más relevantes en la banca minorista. Según Toledo Concha (2022) los préstamos personales son riesgosos debido a existe dificultad de seguimiento, plazos grandes e inclusive garantías físicas para poder obtenerlos. Para una institución financiera, entender el comportamiento de pago de los clientes que

solicitan este tipo de crédito no es solo conveniente, si no es una necesidad operativa y estratégica. La pandemia del COVID-19 puso en evidencia la fragilidad de esa comprensión. El desempleo y los gastos de salud dispararon la morosidad en casi todos los productos bancarios. Un informe de Equifax, Gestión (2021) señala que la morosidad de los préstamos personales creció 3 puntos porcentuales; junto con las tarjetas de crédito, concentraron el 37% de la morosidad total del país. En julio de 2020, la deuda acumulada de los morosos en este segmento alcanzó los S/9.235 millones, esto se puede apreciar en la Figura 1.

Figura 1.

Evolución de deuda y el número de morosos en año 2019-2020 en el Perú



Nota. Fuente: Gestión (2021)

El problema no es peruano ni coyuntural. En Latinoamérica, los préstamos personales suelen ser más atractivos que las tarjetas de crédito: tasas de interés más bajas, montos más altos y plazos más flexibles. Infobae (2022) señala que, en la región, son junto con los depósitos a plazo los productos más demandados. Eso convierte su monitoreo en una prioridad para cualquier banco que quiera evaluar nuevos créditos, reprogramar deudas o ampliar la oferta de productos sin exponerse a pérdidas innecesarias. En el Perú, durante la pandemia, los préstamos

personales ocuparon el tercer lugar en morosidad y representaron el 19% de la deuda crediticia total (figura 2).

Figura 2.

Distribución de cantidad de morosos y deuda financiera por producto crediticio



Nota. Fuente: Infobae (2022)

La falta de monitoreo sistemático del comportamiento de pago tiene consecuencias concretas. Un banco que no anticipa el impago pierde dinero, deteriora su rentabilidad y compromete su solvencia. También se arriesga a reducir su cartera de crédito, lo que limita su capacidad de otorgar nuevos préstamos y de crecer. Si los clientes incumplen, la respuesta institucional suele ser subir tasas o reducir montos, medidas que dañan la relación con los clientes y erosionan la reputación en el mercado.

El monitoreo constante, en cambio, permite identificar patrones de pago, detectar señales tempranas de deterioro y ofrecer soluciones antes de que el cliente entre en mora. También ayuda a ajustar políticas de crédito de forma proactiva.

Ahí es donde el machine learning entra en juego. Como señala Giraldo y Marín (2021) los bancos están expuestos permanentemente al riesgo crediticio, y una de las formas más

eficaces de medir la probabilidad de impago es mediante algoritmos de ML. Aplicar estas técnicas al seguimiento de la cartera de préstamos personales permite evaluar a los clientes con productos vigentes usando datos históricos. El resultado es una estimación, con un horizonte temporal definido, de la probabilidad de que un cliente caiga en default. Esa estimación puede servir como filtro para renovar productos, como herramienta de seguimiento de la cartera activa y, en general, como un insumo transversal para la gestión del riesgo.

1.3 Formulación del problema

1.3.1 Problema General

¿De qué manera se puede realizar un modelo de Machine Learning basado en el algoritmo XGBOOST para medir la probabilidad de impago de clientes en una entidad bancaria,2026?

1.3.2 Problema Específicos

- ¿De qué manera se puede construir un dataset inicial que contenga variables de carácter crediticio, social y comportamental sobre clientes en una entidad bancaria,2026?
- ¿Cómo se puede realizar el preprocesamiento del dataset inicial que va a servir para medir la probabilidad de impago de clientes en una entidad bancaria,2026?
- ¿De qué manera se puede obtener el mejor modelo para el dataset tratado e imputado para medir la probabilidad de impago de clientes en la cartera de Préstamos Personales,2026?
- ¿Como se puede implementar un seguimiento de performance, estabilidad e interpretabilidad para el Modelo que mide la probabilidad de impago de clientes en una entidad bancaria, 2026?

1.4 Objetivos de la investigación

1.4.1 Objetivo General

Desarrollar un Modelo de Machine Learning usando el algoritmo XGBOOST para medir la probabilidad de impago de clientes en una entidad bancaria, 2026

1.4.2 Objetivos Específicos

- Identificar las variables de carácter crediticio, social y comportabilidad para la construcción del dataset inicial que sirva para medir la probabilidad de impago de clientes en una entidad bancaria, 2026.
- Preprocesar el dataset inicial que va a servir para medir la probabilidad de impago de clientes en una entidad bancaria, 2026.
- Entrenar el mejor Modelo de Machine Learning para medir probabilidad de impago de clientes en una entidad bancaria, 2026.
- Implementar el monitoreo del Modelo que mide la probabilidad de impago de clientes en una entidad bancaria, 2026.

1.5 Justificación de la investigación

La necesidad de mejorar la gestión del riesgo crediticio en préstamos personales justifica esta investigación. Los bancos peruanos enfrentan un escenario donde la morosidad en este producto creció 3 puntos porcentuales durante la pandemia (Gestión, 2021) y representa una parte significativa de la cartera total. Un modelo predictivo basado en machine learning puede reducir pérdidas por impagos, optimizar la rentabilidad y, al mismo tiempo, permitir ofrecer productos más ajustados al perfil de cada cliente. Eso no solo protege al banco, también mejora la experiencia del cliente: menos rechazos injustificados, más ofertas pertinentes. En

un mercado competitivo, la capacidad de anticipar el comportamiento de pago se convierte en una ventaja estratégica.

1.5.1 Justificación Teórica

Desde el punto de vista académico, esta investigación aporta evidencia empírica sobre la aplicabilidad de algoritmos de boosting (XGBoost, LightGBM) en el contexto financiero peruano. Aunque la literatura internacional ha demostrado su superioridad predictiva (Liu et al., 2023; Aruleba y Sun, 2024), aún existen vacíos en cuanto a su implementación en entornos regulados con alta desproporción de clases y restricciones de interpretabilidad (Hermitaño Castro, 2022). Este estudio contribuye a cerrar esa brecha al proponer un modelo que integra rendimiento, explicabilidad (SHAP) y monitoreo continúo alineado con la normativa SBS. Además, al utilizar datos de una entidad bancaria real, genera conocimiento situado que puede ser transferido a otras instituciones de la región. Desde una perspectiva social, un modelo más justo y transparente reduce el sesgo algorítmico y favorece la inclusión financiera de segmentos con historial crediticio limitado (Zegarra Valdivia y Gil Aguilar, 2025).

1.5.2 Justificación Práctica

La aplicación de machine learning al seguimiento de la cartera de préstamos personales ofrece ventajas operativas concretas. Primero, permite procesar grandes volúmenes de datos transaccionales sin pérdida de calidad, algo que los modelos estadísticos tradicionales no logran con la misma eficiencia (de Lange et al., 2022). Segundo, la metodología es escalable: una vez entrenado, el modelo puede actualizarse periódicamente con nuevos datos, adaptándose a cambios en el comportamiento de los clientes. Tercero, la curva de aprendizaje y los costos de implementación son razonables para una entidad financiera mediana; no requiere infraestructura excesiva ni equipos de científicos de datos de gran tamaño. En el caso peruano, donde la SBS exige validación y monitoreo continuo (Resolución S.B.S. N° 00053-2023), un

modelo explicable como el que se propone aquí facilita el cumplimiento normativo sin sacrificar precisión. En definitiva, esta investigación ofrece una hoja de ruta práctica para que los bancos adopten inteligencia artificial en la gestión del riesgo crediticio de forma efectiva, transparente y sostenible.

1.6 Alcance

Según Hernández Sampieri & Fernandez-Collado (2014), se define al alcance como la profundidad que tendra el estudio, se establece los limites conceptuales y metodológicos del proyecto. Por lo tanto se indica que esta investigación se centra en el desarrollo de un modelo de machine learning para estimar la probabilidad de impago de clientes en una entidad bancaria peruana. El alcance cubre todo el ciclo de vida del modelo: desde el entrenamiento hasta el despliegue y el monitoreo continuo. El marco de desarrollo se alinea con la normativa de gestión de riesgos de modelos de la Superintendencia de Banca y Seguros (SBS).

Se probarán múltiples enfoques algorítmicos: Regresión Logística, Árboles de Decisión, Random Forest y algoritmos basados en boosting, entre otros. Todos los modelos se entrenarán sobre un conjunto de datos que pasará por un proceso de preprocesamiento exhaustivo, con filtros estadísticos y de negocio para garantizar la calidad y relevancia de la información.

El desempeño de cada modelo se evaluará con métricas estándar: área bajo la curva ROC (AUC) y coeficiente de Gini, complementadas con indicadores auxiliares para monitorear la salud del modelo final. Una vez seleccionado el mejor candidato, se aplicará un proceso de optimización para refinar sus métricas. Luego se desplegará en entornos de nube mediante pipelines automatizados, lo que permitirá generar puntuaciones (scoring) y calcular métricas de monitoreo de forma continua.

Como entregable final, se desarrollará una plataforma web que integre todos estos componentes. Los analistas técnicos y de negocio podrán acceder a los resultados del modelo, validar su comportamiento y entender su aplicación sobre la cartera de clientes. La plataforma incluirá un tablero de control con reportes 360° sobre el rendimiento y la estabilidad del modelo, además de un agente especializado basado en inteligencia artificial generativa (Gen AI). Este agente interpretará las métricas, facilitará la detección temprana de deterioros e identificará oportunidades de mejora en la gestión o en futuros reentrenamientos. Con esto, el ciclo de vida del modelo queda cubierto desde su concepción hasta su operación y evolución continua.

1.7 Limitaciones de la investigación

Según Hernández Sampieri & Fernandez-Collado (2014), las limitaciones son una parte escencial al plantear el proyecto, ya que reflejan las condiciones que pueden restarle confiabilidad a un estudio. Por ello resulta necesario explicar las principales condiciones limitantes aperturadas a nivel espacial, temporal y social.

1.7.1 Espacial

El modelo se entrenará con datos de una sola entidad bancaria peruana. Esto limita la aplicabilidad de los resultados a otras instituciones o países. Las distribuciones de las variables explicativas como ingresos declarados, antigüedad laboral, historial de pagos pueden variar significativamente entre regiones. Un cliente de Lima Metropolitana no tiene el mismo perfil que uno de Arequipa o de la selva, y esas diferencias no siempre están representadas en los datos disponibles. Si el modelo se aplicara a otra cartera sin ajustes, su capacidad predictiva podría deteriorarse. Esta es una restricción inherente a cualquier estudio basado en una única fuente de datos.

1.7.2 Temporal

Los modelos de machine learning entrenados con datos históricos son sensibles a cambios en el entorno económico. La muestra utilizada cubre un período que incluye la pandemia del COVID-19, un evento atípico que distorsionó los patrones de pago. Eso es valioso para capturar comportamientos extremos, pero también introduce ruido: las relaciones entre variables que el modelo aprendió en ese contexto pueden no mantenerse en un escenario de estabilidad económica. La normativa SBS (Resolución N° 00053-2023) exige monitoreo continuo precisamente por esta razón: un modelo que hoy funciona bien puede degradarse mañana si el perfil de los clientes o las condiciones del mercado cambian. Esta investigación aborda el monitoreo, pero no implementa un sistema de reentrenamiento automático, lo que deja abierta la necesidad de actualizaciones periódicas.

1.7.3 Social

Los datos históricos de crédito contienen sesgos. No porque el banco quiera discriminar, sino porque las decisiones de aprobación del pasado reflejan prácticas que hoy consideraríamos injustas: menor acceso al crédito para mujeres, para poblaciones de ciertos distritos, o para personas con empleos informales. Un modelo entrenado con esos datos puede perpetuar esas desigualdades, aunque no sea su intención. Esta investigación incorpora herramientas de explicabilidad como SHAP para detectar sesgos, pero no garantiza su eliminación completa. La equidad algorítmica es un campo en desarrollo, y corregir estos sesgos requiere decisiones de negocio que van más allá de la optimización de una métrica predictiva.

1.8 Identificación de Variables

1.8.1 Variable Independiente (X)

- Modelo de Machine Learning basado en el algoritmo XGBOOST.

1.8.2 Variables predictoras

- 48 características/atributos de los clientes utilizados como variables de entrada del modelo.

1.8.3 Variable Dependiente (y)

- Probabilidad de impago (output del modelo).

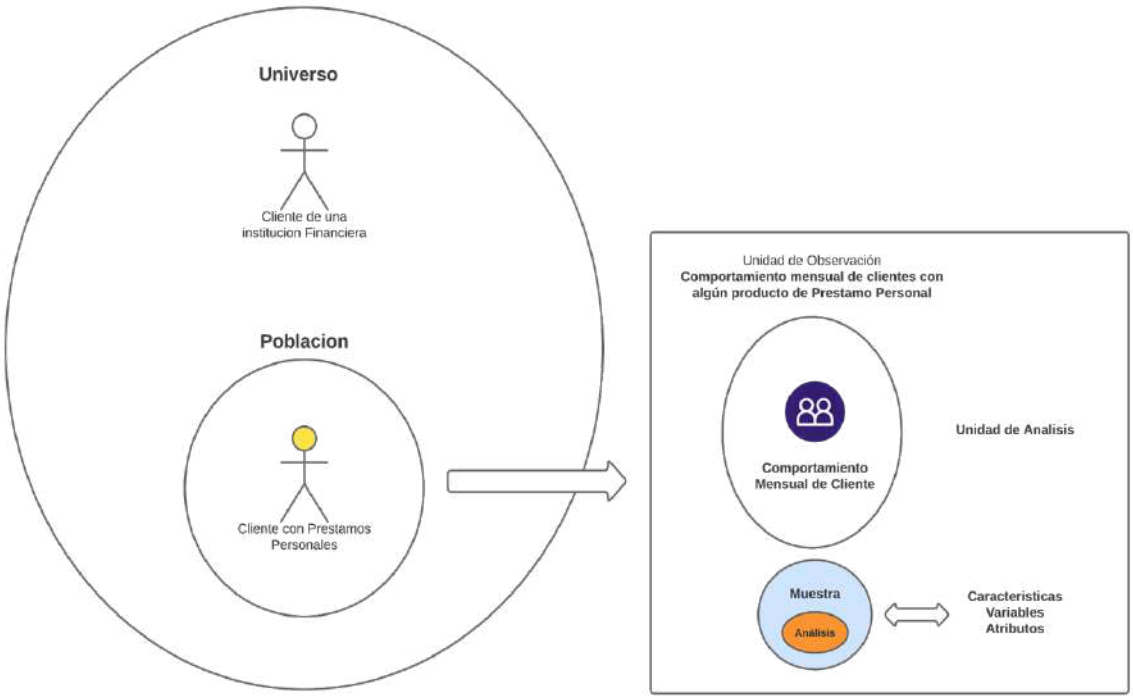
1.9 Universo y muestra

1.9.1 Universo

Para Ñaupas Paitán et al. (2014), se define como universo o población al conjunto de individuos totales que en la práctica pueden no estar completamente disponible para el investigador. En esta investigación el universo está compuesto por todas las personas que han adquirido un producto de la cartera de préstamos personales en la entidad bancaria. Estas personas tienen un crédito vigente o histórico con la institución y, por tanto, una deuda que debe ser cancelada en un plazo determinado. La unidad de análisis observada, como se detalla en la Figura 3, es el comportamiento mensual de los clientes de la cartera de préstamos personales. A través de variables demográficas, socioeconómicas, de comportamiento crediticio, de factor de compra y otras, el modelo buscará predecir la probabilidad de que un cliente caiga en default.

Figura 3.

Diagrama general del Universo-Unidad de Observación de Clientes de la entiedad bancaria



Nota. Fuente: Elaboración propia

1.9.2 Muestra

Hernández Sampieri & Fernandez-Collado (2014), indican que la muestra es un subgrupo de nuestra pobacion de interes, esta tiene que definirse y delimitarse con precision. Por razones prácticas, se extraerá una muestra representativa de la población de clientes de préstamos personales para entrenar el modelo. Se dispone de información desde enero de 2015 hasta mayo de 2026, pero para este estudio se seleccionarán únicamente los periodos comprendidos entre junio de 2021 y octubre de 2022. Esta elección es intencional: incluye tanto meses de pandemia COVID-19 como periodos más estables, lo que añade variabilidad y cierto grado de incertidumbre al modelo, mejorando su capacidad de generalización. Sobre esa base temporal, se aplicarán los siguientes filtros de negocio y estadísticos, con el fin de aumentar la calidad de nuestra muestra:

- Clientes con al menos 3 meses de comportamiento observado.
- Clientes sin garantía hipotecaria asociada al préstamo (este tipo de garantías suele modificar el comportamiento de pago).

- Clientes con 0 días de atraso en el mes de observación (es decir, que estaban al día en ese momento).
- Clientes que no hayan sido castigados por morosidad en otros productos de la entidad.

Con estos criterios, la base inicial queda compuesta por un total estimado de 2,235,642 registros. La Figura 4 muestra el flujo completo de depuración para llegar a la muestra de la investigación.

Figura 4.

Flujo para la obtención de la muestra



Nota. Fuente: Elaboración propia

1.10 Metodología y Diseño de la investigación

1.10.1 Tipo de Investigación

La investigación es de tipo aplicada y predictiva. Es aplicada porque, de acuerdo con Hernández Sampieri & Fernandez-Collado (2014), el propósito fundamental de este tipo de investigación es resolver problemas prácticos en lugar de limitarse únicamente a producir conocimiento básico o teórico. En esta misma línea, Ñaupas Paitán et al. (2014) señalan que la investigación aplicada (o tecnológica) está orientada a resolver objetivamente los problemas de las actividades de servicios y de la vida productiva de la sociedad.

El estudio se enmarca exactamente en este tipo porque busca solucionar un problema concreto del sector bancario: la estimación del riesgo de impago en préstamos personales mediante la implementación de soluciones tecnológicas operativas. Como señalan los autores, no se trata de un ejercicio puramente teórico, sino de transformar la realidad y optimizar sistemas. Por eso, el producto final será un modelo desplegable en producción.

Es predictiva porque, según Ñaupas Paitán et al. (2014), la investigación con fines de predicción busca "futurizar" o prever el comportamiento que tendrán los objetos, eventos o procesos, sirviendo para prevenir o advertir sobre situaciones futuras. Así, la investigación es predictiva porque su objetivo central es desarrollar algoritmos de machine learning que generen pronósticos individualizados de comportamiento futuro. Esto permite anticipar escenarios de incumplimiento antes de que ocurran.

1.10.2 Nivel de Investigación

El nivel de investigación es explicativo, con un componente predictivo. Es explicativo porque, según Hernández Sampieri & Fernandez-Collado (2014), este alcance va más allá de la simple descripción de fenómenos o del establecimiento de relaciones. De este modo, el estudio busca explicar relaciones entre un conjunto de variables predictoras como historial crediticio, comportamiento transaccional, características socioeconómicas y la variable dependiente (probabilidad de impago), utilizando técnicas de aprendizaje supervisado.

El componente predictivo se sustenta en que la utilidad principal de asociar variables es saber cómo se comportará un concepto al conocer el comportamiento de otras variables vinculadas. Al conocer la magnitud de esta asociación, se tienen las bases para intentar predecir el valor aproximado que tendrá un caso. Esto permite generar puntuaciones de riesgo que anticipan el incumplimiento de pago, superando el alcance de un estudio meramente descriptivo.

1.10.3 Enfoque de Investigación

El enfoque es cuantitativo. Según Ñaupas Paitán et al. (2014), este enfoque se caracteriza por utilizar métodos y técnicas cuantitativas que tienen que ver con la medición, el uso de magnitudes, el muestreo y el riguroso tratamiento estadístico. Los autores explican que este enfoque utiliza la recolección de datos y el análisis de estos para contestar preguntas de investigación y probar hipótesis, confiando en la medición de variables y en el uso de la estadística descriptiva e inferencial.

En el contexto del estudio, esto se evidencia directamente: el proyecto se basa en la recolección y análisis de datos numéricos estructurados y en la aplicación de algoritmos matemáticos para identificar patrones en grandes volúmenes de información. Al trabajar con muestras representativas de clientes de una institución financiera, los resultados se pueden evaluar y generalizar mediante métricas objetivas como AUC-ROC, Gini, precisión y sensibilidad, lo que permite establecer conclusiones con sólido respaldo empírico y estadístico.

1.10.4 Diseño de Investigación

El diseño es no experimental. Según Hernández Sampieri y Fernández-Collado (2014), la investigación no experimental se realiza sin manipular deliberadamente variables. En este tipo de estudio, el investigador no construye ni provoca intencionadamente ninguna situación o estímulo, sino que se limita a observar los fenómenos tal como se dan en su ambiente y contexto natural para analizarlos.

Bajo este fundamento, en la presente investigación no se manipula intencionalmente ninguna variable independiente. No se asigna aleatoriamente a los clientes a grupos de control y tratamiento, ni se interviene sobre su comportamiento crediticio. En su lugar, se emplea un conjunto de datos observacionales históricos de clientes con préstamos personales vigentes entre junio de 2021 y octubre de 2022. El investigador se limita a observar y recolectar la

información disponible en las bases de datos de la entidad bancaria, sin intervenir en las condiciones que generaron los perfiles de pago de los clientes.

1.10.5 Complemento metodológico: Ciencias del diseño (DSR)

Los apartados previos definen el diseño general de esta investigación como cuantitativo, de tipo aplicado, con nivel explicativo-predictivo y diseño no experimental. Pero existe un detalle, la construcción del artefacto tecnológico el modelo predictivo, el pipeline CRISP-ML(Q), la plataforma y el agente no se reduce a observar datos y medir relaciones. Implica diseñar, construir y refinar una solución. Esa lógica de construcción responde a un enfoque distinto, conocido como Ciencias del Diseño (Design Science Research), que complementa al diseño no experimental sin reemplazarlo.

March y Smith (1995) plantean una distinción fundamental en las disciplinas aplicadas que es estar orientada a construir y evaluar artefactos, como constructos, modelos, métodos o instanciaciones que resuelven problemas prácticos aún sin solución. El modelo predictivo desarrollado en este estudio es, justamente, un artefacto de tipo modelo o método, pensado para dar respuesta al problema del riesgo de impago en préstamos personales.

Peffers et al. (2007) formalizan esta lógica constructiva a través de la Metodología de Investigación en Ciencia del Diseño (DSRM), que propone un proceso ordenado en fases: identificación del problema, definición de objetivos, diseño y desarrollo del artefacto, demostración, evaluación y comunicación de resultados. Según los autores, las disciplinas aplicadas como los Sistemas de Información requieren este paradigma porque describir la realidad no basta; es necesario intervenir sobre ella construyendo una solución concreta. El presente estudio adopta, dentro de esta metodología, un enfoque centrado en el problema (problem-centered approach), ya que la necesidad de estimar el riesgo de impago surge directamente de la práctica operativa de la entidad financiera y no de una motivación puramente

teórica. Así, las fases de identificación del problema, definición de objetivos, y diseño y desarrollo del artefacto siguen la lógica iterativa de construcción propia de la DSRM.

1.11 Organización de la tesis

El presente trabajo de tesis está organizado en cinco capítulos. A continuación, se resume el contenido de cada uno.

- Capítulo I: Planteamiento Metodológico. Este capítulo contiene el planteamiento del problema, la justificación, los objetivos y la metodología de la investigación. Aquí se delimita el alcance del estudio, se definen las variables de investigación, se describe el universo y la muestra, y se detalla el diseño metodológico empleado: tipo, nivel, enfoque y diseño de la investigación.
- Capítulo II: Marco Teórico. El capítulo II desarrolla los conceptos y términos fundamentales para la comprensión del tema. Se abordan el riesgo crediticio, los modelos de machine learning supervisado (XGBOOST, LightGBM, Random Forest, entre otros), las métricas de evaluación (AUC-ROC, Gini, KS, matriz de confusión) y la normativa de gestión de riesgos de modelos establecida por la Superintendencia de Banca y Seguros (SBS). Estos conceptos sirven como base teórica a lo largo de todo el documento.
- Capítulo III: Estado del Arte Metodológico. El capítulo III presenta el estado del arte metodológico. Revisa investigaciones previas relacionadas con la predicción de impago en el sector bancario. Se exploran diversos enfoques algorítmicos, técnicas de preprocesamiento de datos, ingeniería de variables y estrategias de validación de modelos. También se revisan experiencias previas en el despliegue de modelos en producción y su monitoreo continuo. Todo ello permite contextualizar y fundamentar la solución propuesta.

- Capítulo IV: Implementación y Desarrollo de la Solución Propuesta. El capítulo IV se dedica a la implementación y desarrollo de la solución propuesta. Detalla todo el ciclo de vida del modelo: la construcción del dataset inicial, la definición de la variable objetivo mediante análisis de Roll Rates, la aplicación de filtros de inclusión y exclusión, el preprocesamiento y limpieza de datos, el análisis bivariado y transformación de variables, el tratamiento de valores nulos y outliers, y el entrenamiento y validación de múltiples algoritmos candidatos. También se explica la optimización bayesiana de hiperparámetros, la evaluación de performance del modelo final, la transformación a score crediticio, el análisis de interpretabilidad mediante SHAP, el diseño de la arquitectura en la nube para el monitoreo continuo y el despliegue de la plataforma web Model Risk Management Suite con su agente especializado basado en inteligencia artificial generativa. Además conforme se va desarrollando la implementación de la solución se van presentando 4 artefactos de software (Documento de análisis de software, documento de arquitectura de software, documentación de diseño de interfaz y experiencia de usuario, documento de pruebas de software).
- Capítulo V: Discusión, Conclusiones y Recomendaciones. Por último, el capítulo V engloba la discusión de los resultados obtenidos, contrastando los hallazgos con investigaciones previas. También incluye las conclusiones generales del trabajo, las limitaciones encontradas y las recomendaciones para futuras investigaciones. Se proponen líneas de mejora en la gestión de modelos, posibles reentrenamientos y la evolución del agente inteligente para la interpretación automatizada de métricas.

CAPÍTULO II: MARCO TEÓRICO

En este capítulo se detalla el glosario de términos que se utilizan a lo largo del trabajo. Las definiciones de cada concepto proceden de artículos científicos y fuentes confiables.

2.1 Riesgo crediticio

El riesgo crediticio se define como la posibilidad de que un prestatario no cumpla con sus obligaciones financieras, lo que puede generar pérdidas económicas para bancos y otras instituciones financieras (Alizade, 2025). Este tipo de riesgo está presente en diversas operaciones, como préstamos, bonos, crédito comercial y otras formas de financiamiento (Scott et al., 2024). Por eso, su adecuada gestión resulta clave para garantizar la estabilidad y solvencia de las entidades financieras.

2.1.1 Préstamos Personales

En el contexto de los préstamos personales, el riesgo de incumplimiento puede surgir desde la etapa de evaluación inicial del cliente y mantenerse durante toda la vida del crédito (Calin et al., 2023). Los préstamos personales son un tipo de producto bancario a largo plazo y su monto puede oscilar entre cantidades pequeñas y cantidades grandes. Cuando el deudor deja de pagar según lo acordado, se produce la morosidad o default (Egwa et al., 2022).

2.1.2 Morosidad y Default

La morosidad son los atrasos (días de mora o DM) que presentan os clientes, y el default es la marca que se le da a un cliente cuando este pasa un umbral específico de dm. La evidencia reciente muestra que el riesgo de impago no depende únicamente de las características individuales del prestatario, sino también del entorno económico y político en el que se desenvuelve (Saliba et al., 2023). En el caso de Italia, por ejemplo, se encontró que el desempleo y la expansión del crédito se asociaron con mayores niveles de préstamos

deteriorados (Foglia, 2022). Mientras tanto, variables macroeconómicas favorables tendieron a reducirlos. Esto evidencia la importancia de considerar factores contextuales al momento de modelar el comportamiento de pago de los clientes.

2.1.3 Scoring Crediticio

El scoring crediticio es una herramienta de evaluación que permite identificar qué solicitantes tienen mayor probabilidad de incumplir con sus pagos (Wu et al., 2025). Los estudios recientes coinciden en que los modelos de machine learning suelen superar a los métodos estadísticos tradicionales en términos de precisión predictiva (Mestiri, 2024). Esta ventaja se acentúa cuando se incorporan grandes volúmenes de datos, variables no lineales o información conductual (Abi, 2025). Además, se demuestra que el rendimiento de estos modelos mejora significativamente cuando se corrige el desbalance de clases y se incorporan nuevas fuentes de información, como datos transaccionales, texto o variables no financieras También se ha observado que la inversión en tecnologías financieras (Fintech) puede contribuir a reducir los préstamos en mora y mejorar el desempeño bancario, aunque estos efectos tienden a manifestarse con cierto rezago temporal (H. Wang et al., 2023).

2.2 Machine Learning

El machine learning es una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y mejorar su desempeño sin necesidad de programación explícita para cada regla (Alnuaimi & Albaldawi, 2024). Su desarrollo reciente ha sido acelerado y se asocia tanto con mayores capacidades computacionales como con la expansión de aplicaciones predictivas y analíticas en múltiples campos (Hsu et al., 2023). La evolución del campo ha ampliado su alcance desde tareas de predicción hacia usos de análisis e inferencia, aunque la interpretación causal sigue siendo un tema discutido.

El machine learning es un campo interdisciplinario que combina estadística, informática y otras áreas del conocimiento. La literatura reciente destaca tendencias hacia la inteligencia artificial explicable, responsable y causal, lo que refleja una maduración del campo hacia aplicaciones más confiables y transparentes. Sin embargo, persisten desafíos importantes en cuanto a la calidad de los datos, la capacidad computacional y la implementación técnica en entornos reales (Suresh Dameruppula, 2025).

2.2.1 Aprendizaje Supervisado

El aprendizaje supervisado utiliza datos etiquetados para entrenar modelos que asignan entradas a salidas conocidas. Sus dos problemas principales son la clasificación, cuando la salida es categórica, y la regresión, cuando la salida es continua (Shetty et al., 2022). En el caso de la clasificación, el objetivo es aprender una representación compacta de la relación entre los predictores y las etiquetas, para luego asignar clases a nuevos casos no vistos.

Un conjunto de entrenamiento más grande tiende a mejorar el rendimiento del modelo bajo ciertos supuestos estadísticos. La selección del algoritmo adecuado depende de características de los datos como la heterogeneidad, la precisión requerida, la redundancia de variables y la linealidad de las relaciones. Las aplicaciones comunes del aprendizaje supervisado incluyen reconocimiento de patrones, bioinformática, procesamiento de texto e imágenes, entre otras (Han et al., 2012).

2.2.2 Clasificación Binaria y Balanceo de Clases

La clasificación binaria busca separar observaciones en dos categorías mutuamente excluyentes, como default/no default, fraude/no fraude o mayoría/minoría (Li & Sha, 2024). En problemas del mundo real, el desbalance de clases es un fenómeno frecuente que sesga a los modelos hacia la clase mayoritaria, lo que reduce significativamente el desempeño sobre la clase de interés. Esta situación es particularmente crítica en el sector bancario, donde los casos

de impago (clase minoritaria) son los que se desean detectar con mayor precisión (Husain et al., 2025).

Para corregir este problema, se utilizan diversos enfoques. A nivel de datos, se emplean técnicas como oversampling (sobre-muestreo), undersampling (sub-muestreo) y métodos híbridos como SMOTE-ENN (Wongvorachan et al., 2023). También existen enfoques a nivel de algoritmo que modifican la función de costo o el peso de las clases. La técnica SMOTE (Synthetic Minority Oversampling Technique) genera muestras sintéticas mediante interpolación entre vecinos de la clase minoritaria, lo que permite equilibrar artificialmente el conjunto de datos (Elreedy et al., 2024).

La evidencia reciente muestra que SMOTE-ENN suele mejorar la exactitud, la estabilidad y la capacidad de generalización de los modelos en comparación con SMOTE solo. Sin embargo, el mejor método depende del grado de desbalance presente en los datos: el oversampling simple (ROS) rindió mejor en casos de desbalance moderado, mientras que los métodos híbridos demostraron ser más efectivos en situaciones de desbalance extremo. En el contexto de detección de fraudes, también se ha reportado que SMOTE-ENN contribuye a mejorar la capacidad predictiva de los modelos (Bounab et al., 2024).

2.2.3 Algoritmos de Clasificación

2.2.3.1 Regresión Logística

La regresión logística es una técnica de clasificación utilizada para estimar la probabilidad de un resultado binario a partir de variables independientes. Su formulación conecta la probabilidad del evento con una combinación lineal de predictores mediante la función logit y el uso de razones de momios u odds ratio (Balboa et al., 2024).

La regresión logística sigue siendo un referente en clasificación supervisada por su simplicidad interpretativa y por su capacidad para modelar probabilidades de pertenencia a una

clase (L.-F. Wang et al., 2025). En aplicaciones recientes, se ha utilizado como línea base frente a modelos más complejos, y en algunos contextos conserva un buen desempeño predictivo, especialmente en términos de discriminación medida por AUC. Su principal ventaja radica en la transparencia de sus coeficientes, lo que facilita la explicación de los resultados ante equipos de negocio y entes reguladores.

2.2.3.2. Árboles de Decisión

Los árboles de decisión son modelos de clasificación que construyen una estructura jerárquica de decisiones a partir de divisiones sucesivas de los datos. Cada trayectoria desde la raíz hasta una hoja representa una regla de decisión, lo que favorece su interpretación en problemas con variables numéricas y categóricas (Balboa et al., 2024).

Los árboles de decisión ocupan un lugar central en la clasificación porque generan reglas comprensibles para el usuario y admiten distintos tipos de variables de entrada (Lee et al., 2022). Además, sirven como base estructural para varios métodos de ensamble posteriores, tanto de bagging como de boosting (Boldini et al., 2023). Su principal limitación es la tendencia al sobreajuste cuando no se podan adecuadamente, aunque su simplicidad los convierte en una herramienta valiosa para explorar relaciones en los datos (Zhu et al., 2025).

2.2.3.3 Random Forest

Random Forest es un método de ensamble basado en la construcción de múltiples árboles de decisión, donde la clase final se define por votación entre los modelos generados (Shinagam et al., 2024). Su funcionamiento se basa en entrenar muchos árboles sobre subconjuntos aleatorios de los datos y combinar sus salidas por votación, lo que reduce la variabilidad de los resultados y suele mejorar el rendimiento frente a modelos individuales.

Este algoritmo es uno de los métodos de ensamble más difundidos dentro del aprendizaje supervisado (Nithya et al., 2025). Su capacidad para manejar grandes volúmenes

de datos, relaciones no lineales e interacciones entre variables lo convierte en una opción robusta para problemas de clasificación en el sector bancario. Adicionalmente, Random Forest proporciona medidas de importancia de variables que permiten identificar los predictores más relevantes para el modelo.

2.2.3.4 AdaBoost

AdaBoost es un algoritmo de boosting que combina clasificadores débiles en un clasificador fuerte, reasignando mayor peso a los casos mal clasificados en iteraciones sucesivas (Ramakrishna et al., 2023). Fue uno de los primeros algoritmos de boosting exitosos y se caracteriza por convertir clasificadores débiles con frecuencia árboles pequeños o decisión stumpsen un clasificador más robusto

Su lógica consiste en aumentar el peso de las observaciones mal clasificadas para que los siguientes modelos concentren mayor atención en esos casos difíciles (Mienye & Sun, 2022a). Este enfoque iterativo permite que AdaBoost mejore progresivamente el rendimiento del modelo, aunque puede ser sensible a datos ruidosos y valores atípicos que reciben pesos elevados en etapas avanzadas del entrenamiento.

2.2.3.5 CatBoost

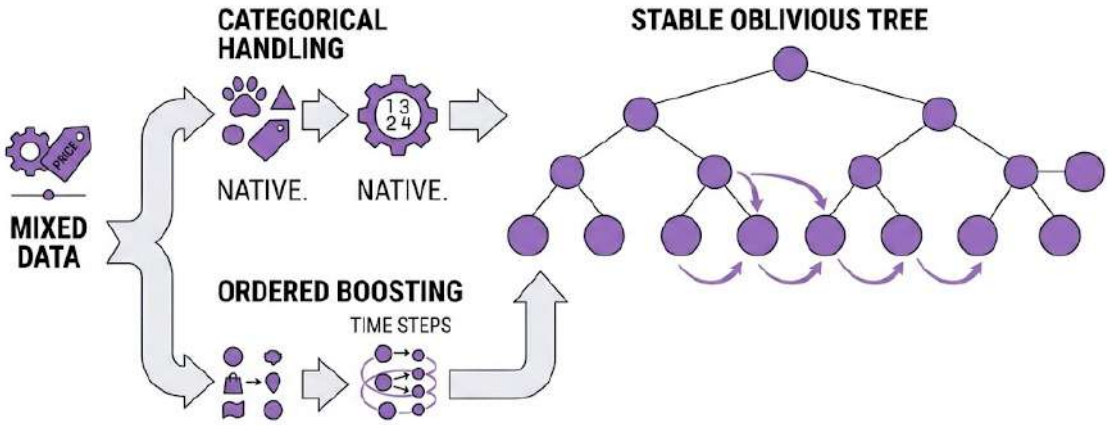
CatBoost es una variante de gradient boosting diseñada para tratar variables categóricas y reducir el problema de predicción shift mediante ordered boosting (Arini et al., 2026). Sobresale porque fue diseñado para procesar eficazmente variables categóricas y para mitigar el prediction shift, un fenómeno que afecta la capacidad de generalización de los modelos (Boldini et al., 2023).

Además, CatBoost utiliza árboles simétricos, lo que reduce el tiempo de predicción y el riesgo de sobreajuste en comparación con implementaciones basadas en árboles asimétricos. Esta característica, sumada a su capacidad para manejar datos categóricos sin necesidad de

codificación previa, lo convierte en una opción atractiva para problemas bancarios donde abundan variables de este tipo. En a Figura 5 se observa cómo se puebla el algoritmo de CatBoost a través de un cálculo de los residuos de manera secuencial para evitar que la información del futuro "contamine" el entrenamiento actual, resultando en modelos más robustos y estables, especialmente con datos mixto

Figura 5.

Construccion de Arbol del Algoritmo CatBoost



Nota. Fuente: Elaboración propia

2.2.3.6 LightGBM

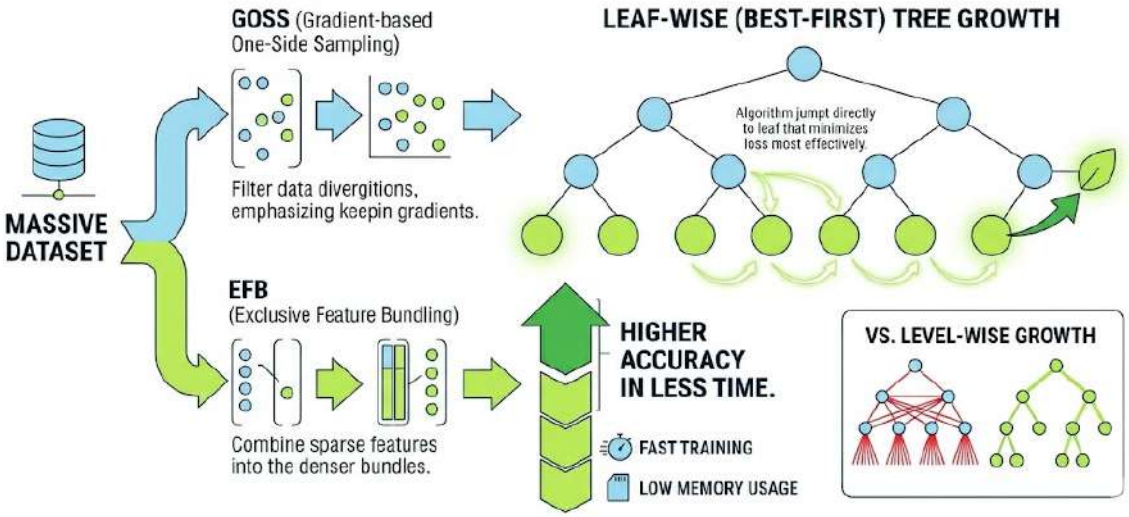
LightGBM es una implementación de gradient boosting reconocida como parte de la familia de algoritmos que mejora el rendimiento predictivo mediante optimizaciones específicas del entrenamiento (Villar & De Andrade, 2024). Se caracteriza por su eficiencia computacional y su capacidad para manejar grandes conjuntos de datos, lo que lo hace particularmente útil en entornos bancarios con volúmenes masivos de información.

LightGBM introduce técnicas como el muestreo basado en gradientes (Gradient-based One-Side Sampling) y el agrupamiento de características exclusivas (Exclusive Feature Bundling), que reducen significativamente el tiempo de entrenamiento sin sacrificar la

precisión predictiva. En la Figura 6 se observa como en lugar de construir el árbol nivel por nivel, LightGBM identifica la hoja específica con el mayor potencial de reducción de pérdida y la divide inmediatamente, lo que resulta en estructuras asimétricas y una convergencia mucho más rápida

Figura 6.

Construccion de Arbol del Algoritmo LightGBM



Nota. Fuente: Elaboración propia

2.2.3.7 XGBOOST

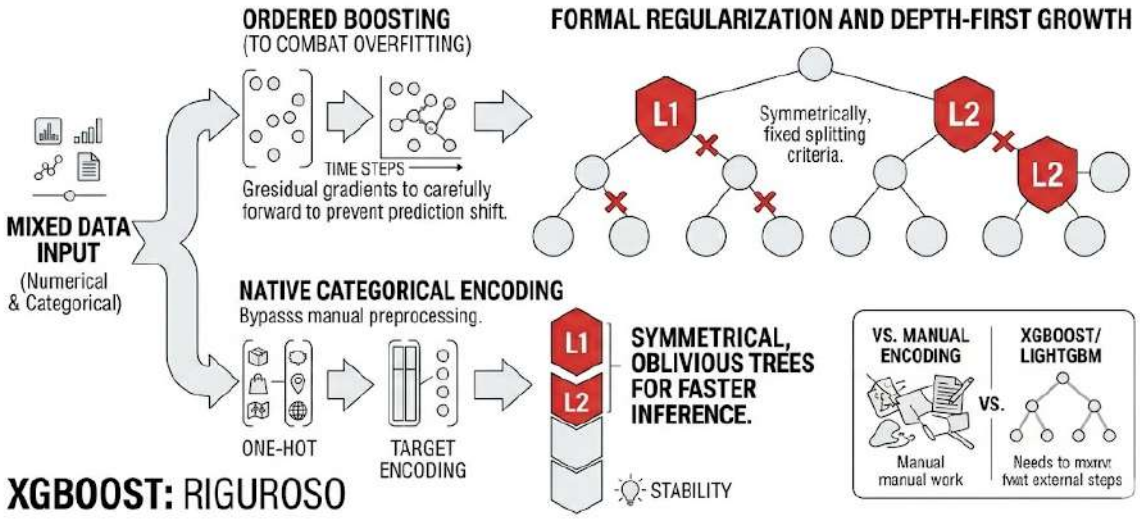
XGBoost es un sistema escalable de árboles potenciados que aprende secuencialmente del error del árbol previo para mejorar el desempeño de clasificación (Campanioni et al., 2024). Es descrito como un algoritmo escalable y altamente preciso para tareas de clasificación y regresión, con capacidad para manejar datos desbalanceados mediante parámetros de regularización y ponderación de clases.

XGBoost incorpora técnicas de regularización L1 y L2 que reducen el sobreajuste y mejoran la capacidad de generalización del modelo (Mienye & Sun, 2022b) . Su versatilidad, velocidad de entrenamiento y rendimiento predictivo lo han convertido en uno de los

algoritmos más utilizados en competencias de machine learning y en aplicaciones bancarias de scoring crediticio. En la Figura 7 se observa el crecimiento Level-wise (por niveles), donde el árbol crece de manera uniforme en profundidad antes de pasar al siguiente nivel, a diferencia de LightGBM

Figura 7.

Construccion de Arbol del Algoritmo XGBoost



Nota. Fuente: Elaboración propia

2.2.4 Métricas de Evaluación

2.2.4.1 Curva ROC y AUC

La curva ROC (Receiver Operating Characteristic) representa el comportamiento de un clasificador binario a través de todos los umbrales posibles. Para ello, grafica la sensibilidad frente a la tasa de falsos positivos (Rainio et al., 2024). Por eso, la ROC no depende de un único punto de corte, sino del recorrido completo de desempeño del modelo sobre distintos umbrales. El área bajo la curva ROC, conocida como AUC, resume ese comportamiento en un solo valor entre 0 y 1 (Yang & Ying, 2023). Además, AUC puede interpretarse como la probabilidad de

que un caso positivo seleccionado al azar reciba una puntuación mayor que un caso negativo también elegido al azar. En contextos con desbalance de clases, varios trabajos sostienen que AUC suele ser más informativa que la exactitud, porque evalúa la capacidad discriminativa del modelo más allá de una sola tasa global de aciertos. Sin embargo, la evidencia reciente también subraya límites importantes. AUC es una medida global y puede ocultar debilidades localizadas en subgrupos específicos de la población evaluada (Sudjianto & Liu, 2025). Asimismo, un valor alto de AUC no garantiza, por sí mismo, un equilibrio adecuado entre sensibilidad y especificidad en un umbral concreto, ni informa directamente sobre precisión o valor predictivo negativo.

2.2.4.2 Coeficiente de Gini

El coeficiente de Gini se relaciona directamente con la curva ROC y el AUC en problemas de clasificación binaria. En particular, la literatura reciente recoge la relación que sigue la ecuación 1.

Ecuación 1:

Calculo del Gini

$$GINI = 2\ x\ AUC - 1 \quad (1)$$

Donde:

- AUC = Area bajo al curva ROC

Ambas medidas expresan la capacidad de discriminación del clasificador en escalas distintas (Dembélé, 2025). Desde esta perspectiva, el Gini transforma el desempeño resumido por AUC a una escala centrada en la separación entre clases. Esto hace que el coeficiente de Gini sea útil cuando se desea expresar de forma compacta cuánto se aparta el modelo de un desempeño aleatorio. No obstante, la evidencia proporcionada en los trabajos recientes

incluidos aquí es más escasa para Gini que para ROC/AUC, y casi siempre aparece como una medida derivada o vinculada a AUC, más que como una métrica desarrollada de forma independiente.

2.2.4.3 Estadístico de Kolmogorov–Smirnov (KS)

En el conjunto de artículos proporcionados, el estadístico Kolmogorov-Smirnov (KS) no aparece desarrollado con el mismo detalle que ROC/AUC, sensibilidad, especificidad o matriz de confusión. Por ello, con base estricta en estas fuentes, solo puede afirmarse que forma parte del ecosistema amplio de métricas de evaluación discutidas en revisiones sobre instrumentos de desempeño para clasificación binaria (Canbek et al., 2022). No obstante, en la práctica bancaria, el KS mide la máxima distancia entre las funciones de distribución acumulada de las probabilidades predichas para los grupos positivos y negativos, y valores superiores a 0.4 suelen considerarse aceptables para fines de discriminación crediticia.

2.2.4.4 Matriz de Confusión

La matriz de confusión es la estructura básica para resumir la correspondencia entre las clases reales y las predicciones del clasificador en un umbral determinado (Richardson et al., 2024). A partir de sus cuatro celdas (verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos) se derivan métricas como exactitud, precisión, sensibilidad y especificidad. Por ello, una parte importante de la evaluación de clasificadores binarios depende de cómo se interpreten conjuntamente estos conteos y no solo una métrica aislada. La literatura reciente insiste en que reportar la matriz completa mejora la interpretación del desempeño, especialmente en contextos sensibles como la investigación médica o los datos desbalanceados (Lavazza & Morasca, 2023). También se ha señalado que distintas métricas derivadas de la matriz pueden reaccionar de forma diferente ante el desbalance de clases y la

incertidumbre muestral, por lo que la comparación entre modelos no debería descansar en una sola medida.

2.2.4.5 Sensibilidad y Especificidad

La sensibilidad expresa la capacidad del modelo para identificar correctamente los casos positivos, mientras que la especificidad representa su capacidad para reconocer correctamente los casos negativos. Estas dos medidas son centrales en la evaluación binaria porque describen los dos tipos principales de acierto del clasificador y suelen entrar en tensión cuando cambia el umbral de decisión (Chicco & Jurman, 2023). Su importancia es especialmente alta en aplicaciones médicas y de alto riesgo, donde maximizar solo una medida puede degradar de forma inaceptable la otra. Además, un punto clave de la literatura reciente es que la interpretación de sensibilidad y especificidad debe hacerse junto con el umbral de clasificación y, de ser posible, junto con la matriz de confusión completa.

2.2.4.6. Índice de Youden

El índice de Youden sintetiza el desempeño diagnóstico combinando sensibilidad y especificidad en una sola medida derivada del punto de corte seleccionado. En los artículos revisados aparece como una de las métricas habituales para resumir el rendimiento de clasificadores binarios junto con exactitud, F1, kappa y MCC. Además, la literatura sobre relaciones entre métricas indica que el estadístico de Youden también se conoce como Bookmaker Informedness, lo que ayuda a ubicarlo dentro de familias de medidas algebraicamente relacionadas (Lovell et al., 2022). Desde un punto de vista interpretativo, el índice de Youden resulta útil cuando se busca un equilibrio entre detección de positivos y exclusión de negativos. Aun así, como ocurre con otras métricas derivadas de la matriz de confusión, su valor depende del umbral elegido y debe interpretarse en conjunto con otras medidas y con el contexto de aplicación.

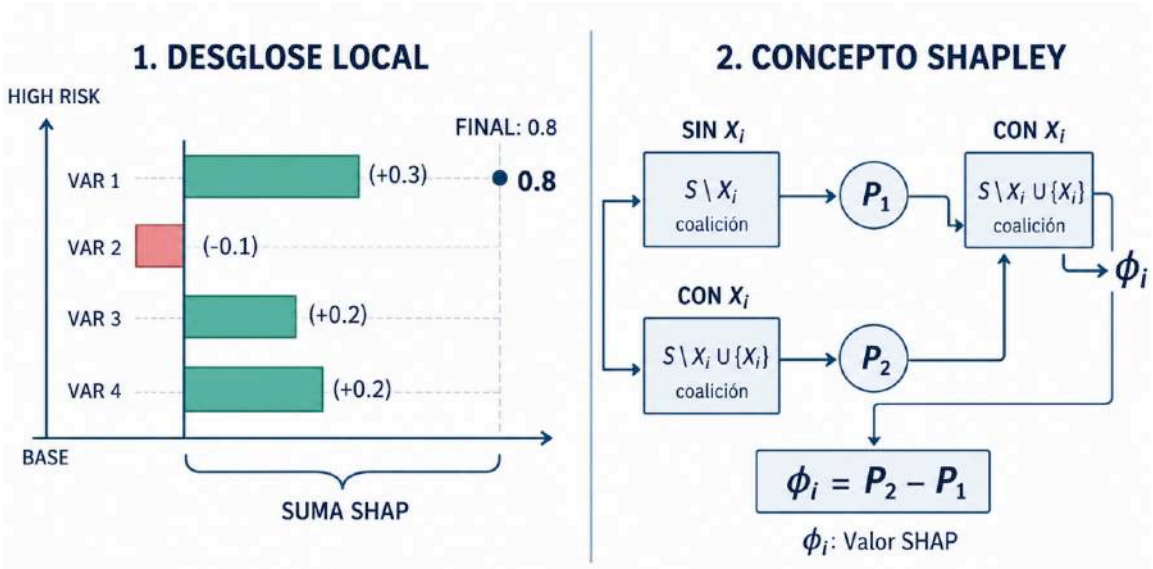
2.2.5 Interpretabilidad de Modelos

2.2.5.1 Importancia de la Interpretabilidad

La interpretabilidad de modelos busca hacer comprensible cómo un modelo de aprendizaje automático produce sus predicciones, con el fin de aumentar la transparencia, la confianza de los usuarios y la utilidad de sus resultados en contextos sensibles como el sector bancario (Salih et al., 2025). Los métodos de explicabilidad post hoc como SHAP y LIME se utilizan ampliamente para traducir modelos complejos a explicaciones más digeribles, especialmente en datos tabulares. Sin embargo, sus resultados dependen del modelo y del conjunto de datos analizado. La interpretación de salidas explicativas exige cautela porque los usuarios pueden sobre confiar en estas herramientas. Factores como la colinealidad entre variables pueden alterar las conclusiones obtenidas. Por ello, se recomienda presentar los resultados junto con sus gráficos y explicar en lenguaje simple sus supuestos, incluida la independencia entre variables. En la Figura 8 se observa como los valores SHAP descomponen una predicción local en contribuciones sumativas de cada variable, calculadas de manera justa promediando su contribución marginal a todas las coaliciones posibles.

Figura 8.

Ejemplo de Descomposición SHAP para la interpretabilidad



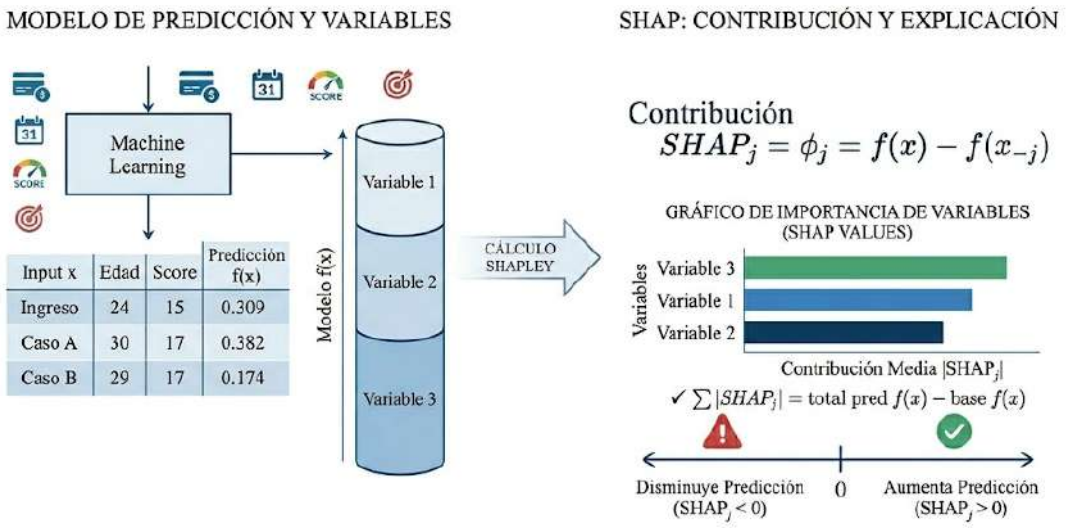
Nota. Fuente: Elaboración propia

2.2.5.2 SHAP (Shapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) es un método de interpretabilidad basado en variables que permite explicar predicciones de modelos supervisados tanto a nivel local como global (Ponce-Bobadilla et al., 2024). Su fundamento proviene de los valores de Shapley, adaptados al aprendizaje automático para cuantificar de forma justa la contribución de cada característica a una predicción individual. SHAP ha ganado amplia adopción por su versatilidad, por la facilidad de implementación en paquetes disponibles y por su utilidad para mejorar la transparencia y la confianza en modelos supervisados. En revisiones recientes, incluso se le describe como un referente importante para las explicaciones locales debido a su base teórica y su aplicabilidad general, aunque existen variantes adaptadas a distintos supuestos y dominios, como el procesamiento de lenguaje natural. Sin embargo, SHAP no debe interpretarse de forma acrítica, porque sus resultados pueden variar según el modelo adoptado. La colinealidad entre variables puede afectar notablemente las explicaciones producidas por SHAP. En la Figura 9 se aprecia de manera general como SHAP permite disgregar la predicción final en base a aportes locales.

Figura 9.

Aporte Local de variables para predicción final en SHAP



Nota. Fuente: Elaboración propia

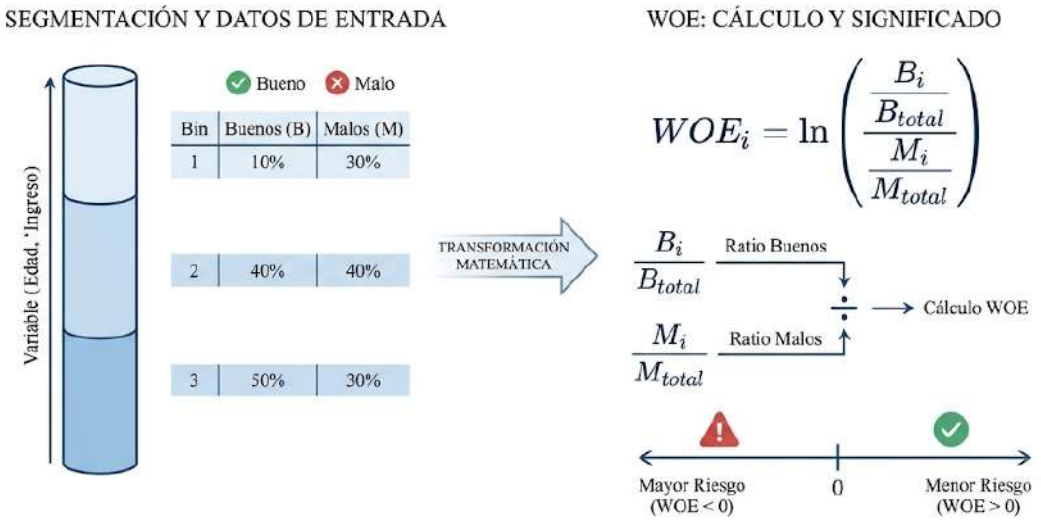
2.2.5.3. Peso de Evidencia (WOE)

El Peso de Evidencia o Weight of Evidence (WOE) es un método de transformación de variables y una técnica estadística bivariada de enfoque log-lineal bayesiano que asigna pesos a categorías o intervalos de una variable según su relación con el evento objetivo (Seitshiro & Govender, 2024). En términos aplicados, WOE expresa la fuerza y dirección de la asociación entre una categoría predictora y el resultado, lo que facilita la interpretación del riesgo relativo o del poder discriminatorio de cada grupo. En regresión logística y credit scoring, WOE se utiliza porque ayuda a representar variables categóricas por niveles de binning que respetan mejor la lógica de los log-odds del modelo. Sin embargo, su impacto sobre el rendimiento predictivo no siempre es positivo: en datos bancarios comerciales, la transformación WOE no mejoró el desempeño del modelo, aunque sí aportó interpretabilidad sobre cada categoría agrupada. Modelos recientes también muestran que WOE puede ofrecer resultados competitivos en otros dominios. En susceptibilidad a deslizamientos de terreno, el modelo

WOE obtuvo tasas de éxito superiores a otros enfoques bivariados comparados y produjo resultados más precisos dentro de ese estudio. En la Figura 10 se observa un ejemplo del cálculo de WOE para una variable.

Figura 10.

Calculo de WOE para una variable continua



Nota. Fuente: Elaboración propia

2.2.5.4 Análisis Bivariado

En este conjunto de estudios, el análisis bivariado aparece principalmente como una lógica de evaluación de la relación entre una variable predictora y una variable objetivo para estimar asociación, capacidad discriminatoria o contribución explicativa. En WOE, esta lógica se operacionaliza asignando pesos según la presencia o ausencia del evento de interés, lo que permite medir la intensidad de la relación entre cada clase de la variable y el resultado observado (Bhandari et al., 2024). Para el marco teórico, puede definirse como un procedimiento exploratorio o modelístico que examina la asociación entre dos variables con

fines de selección, interpretación o predicción. No obstante, conviene señalar que la evidencia disponible aquí no incluye un artículo dedicado específicamente a teorizar el análisis bivariado de forma general, sino aplicaciones concretas dentro de modelos como WOE.

2.2.6 Monitoreo y Gestión de Riesgo de Modelos

2.2.6.1 Monitoreo de Modelos

El monitoreo de modelos es el proceso continuo de supervisar el comportamiento de un modelo una vez desplegado, con el fin de detectar fallas, degradación del rendimiento y cambios en los datos de entrada (Khademi et al., 2023). En entornos de alto impacto, como el sector bancario, este seguimiento no es accesorio, sino un requisito para sostener robustez, seguridad y confiabilidad operacional. La literatura reciente coincide en que el principal reto del monitoreo es identificar diferencias entre la distribución observada durante el desarrollo y la distribución real en producción, ya que esos cambios pueden deteriorar el desempeño incluso después de pruebas satisfactorias (Tyler & McCeney, 2024). También se señala que el monitoreo efectivo no debe limitarse a una sola métrica, porque la robustez depende de analizar simultáneamente el drift de datos, el rendimiento, la explicabilidad y las condiciones del entorno operativo.

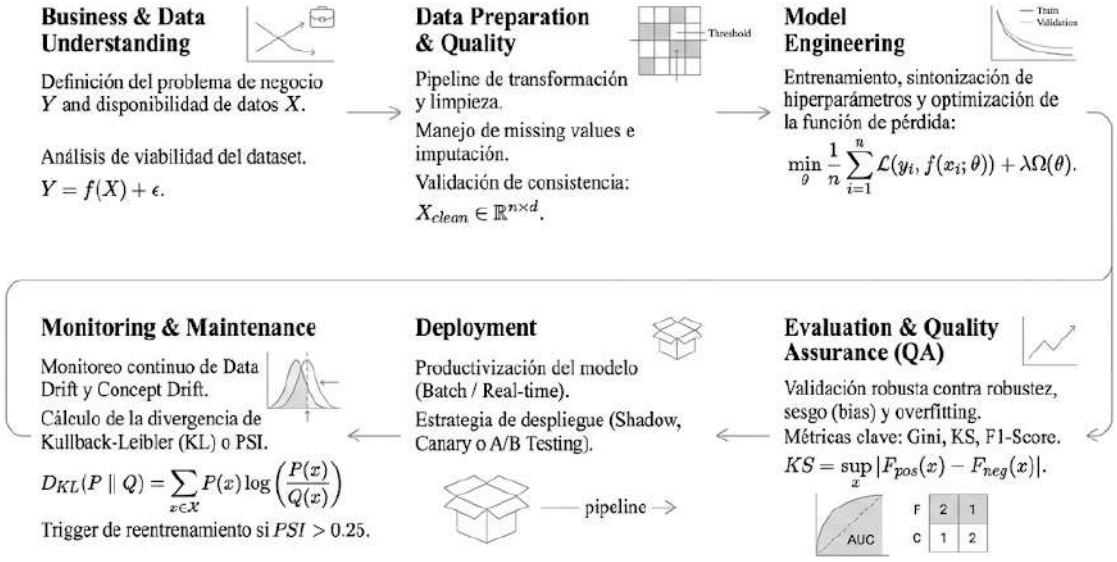
2.2.6.2 CRISP-ML(Q)

CRISP-ML(Q) es un marco metodológico para el ciclo de vida completo de aplicaciones de aprendizaje automático con aseguramiento de calidad, desde la comprensión del negocio y los datos hasta el despliegue, monitoreo y mantenimiento. Su valor teórico radica en que incorpora controles de riesgo y calidad en cada fase, lo que favorece consistencia, confiabilidad y trazabilidad del desarrollo (Magdaong et al., 2024). La evidencia reciente muestra que CRISP-ML(Q) se está utilizando como base para marcos aplicados de mitigación de riesgos, detección de drift, explicabilidad y defensa ante ataques, lo que refuerza su utilidad

para sistemas regulados o sensibles. También se ha destacado que, aunque ofrece estructura y replicabilidad, su aplicación completa puede resultar compleja y debe adaptarse al contexto del proyecto. En la Figura 11 se observa el flujo completo de la metodología CRISP-ML(Q)

Figura 11.

Flujo de metodologia CRISP-ML(Q)



Nota. Fuente: Elaboración propia

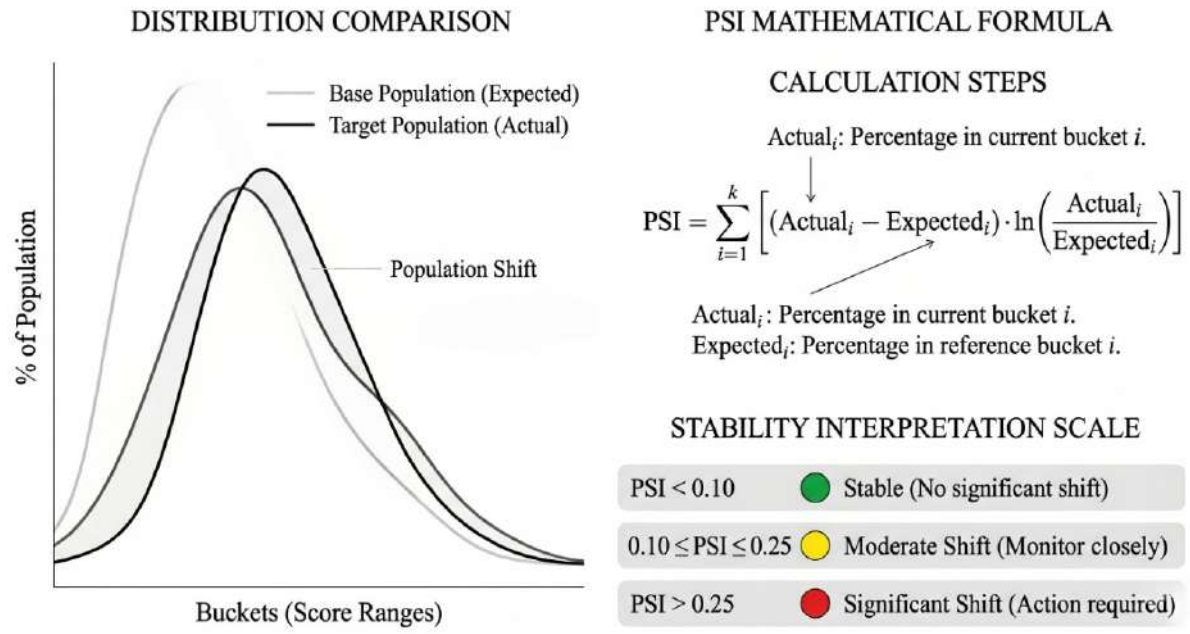
2.2.6.3 Índice de Estabilidad Poblacional (PSI)

El Índice de Estabilidad Poblacional (Population Stability Index, PSI) es una métrica utilizada para detectar y cuantificar cambios en la distribución de variables entre una muestra de referencia y una muestra nueva. Por ello, se utiliza ampliamente en monitoreo de drift de datos y validación continua de modelos en contextos industriales y financieros (Kurian & Allali, 2024). Sin embargo, la evidencia también remarca limitaciones importantes. El PSI requiere variables categóricas o la discretización de variables numéricas, lo que puede producir pérdida de información. Además, su resultado depende del tamaño de los intervalos utilizados.

La interpretación de sus umbrales no siempre está sólidamente discutida en la literatura, por lo que conviene complementarlo con otras métricas de distancia y con indicadores de desempeño del modelo (Agarwal, 2025). En la Figura 12 se observa el cálculo de PSI como ejemplo para una variable así como su respectiva categorización de drift.

Figura 12.

Calculo de PSI y categorizacion de data drift



Nota. Fuente: Elaboración propia

2.2.6.4 Monitoreo de Performance

El monitoreo de performance consiste en verificar si la capacidad predictiva del modelo se mantiene estable a lo largo del tiempo mediante métricas de evaluación apropiadas al problema, como AUC, F1-score, exactitud, recall o métricas rodantes en producción. En riesgo crediticio, este seguimiento es especialmente importante porque un modelo puede exhibir buen rendimiento inicial y aun así generar nuevos riesgos vinculados con estabilidad, interpretabilidad o validación supervisora (Nazir et al., 2025). La literatura reciente muestra

que la evaluación del rendimiento no debe reducirse a la exactitud agregada. También debe considerar la estabilidad de predicciones, la generalización, el sobreajuste y la interpretabilidad, especialmente cuando se utilizan modelos más complejos o ensamblados. En aplicaciones reales, algunos estudios combinan el monitoreo temporal con PSI y métricas móviles para verificar si el desempeño permanece estable en ventanas sucesivas (Nguyen et al., 2025).

2.2.6.5 Gestión de Riesgo de modelos y marco normativo SBS

La Resolución SBS N° 00053-2023 forma parte del marco normativo peruano para la gestión del riesgo de modelos, por lo que en el contexto de esta tesis constituye el referente regulatorio que exige control, validación y seguimiento de los modelos utilizados por entidades supervisadas. Desde la literatura reciente, la gestión del riesgo de modelos se entiende como la identificación y evaluación de riesgos estadísticos, tecnológicos y de conducta o uso, especialmente en modelos de machine learning aplicados a decisiones financieras (Alonso Robisco & Carbó Martínez, 2022). También se advierte que la incertidumbre regulatoria puede frenar la innovación si las organizaciones no cuentan con criterios claros para validar estabilidad, transparencia e interpretación de resultados.

CAPÍTULO III: ESTADO DEL ARTE METODOLÓGICO

3.1 Introducción

La evaluación del riesgo crediticio (credit scoring) es un pilar fundamental en la estabilidad y rentabilidad de los bancos, de ella depende la correcta asignación de recursos y la reducción de pérdidas por mora (default). Durante décadas, los modelos estadísticos lineales especialmente la regresión logística fueron el estándar de la industria. Su sencillez, interpretabilidad y facilidad de implementación los hacían la opción ideal a escoger.

Ese panorama ha cambiado en los últimos años, el crecimiento exponencial del volumen de datos, la aparición de fuentes alternativas de información (textos, redes sociales, comportamientos digitales, etc.) y las crecientes exigencias regulatorias en materia de transparencia y no discriminación (sesgos) han impulsado la adopción de enfoques basados en aprendizaje automático (machine learning, ML) y, más recientemente, también se ha impulsado el uso de modelos de lenguaje de gran escala (LLM).

Esta revisión del estado del arte sintetiza los avances más relevantes publicados a partir de 2021. Los hallazgos se organizan en torno a cuatro ejes tematicos:

- Comparación de modelos avanzados frente a algoritmos tradicionales.
- Uso de datos alternativos.
- Explicabilidad y gobernanza en modelos de IA.
- Uso de LLM e infraestructura.

3.2 Busqueda de papers

Para sintetizar los avances de los 4 ejes anteriormente mencionados, se utilizan los siguientes buscadores: Web of Science (Wos), Scopus, Semantic Scholar y ResearchGate. Se tomo como prioridad artículos y revistas publicadas luego del 2021, ya que es necesario contar con información actualizada sobre el contexto de los modelos crediticios. El procedimiento de

búsqueda se llevó a cabo siguiendo las pautas descritas en la Tabla 1, considerando el título, resumen y palabras clave. Estos términos también se adaptaron para que coincidan con los ejes planteados y se encuentre coincidencias en el motor de búsqueda. Posteriormente, se aplicó un proceso de inclusión y selección de acuerdo con lo mostrado en la Tabla 2.

Tabla 1.

Queries y filtros usados en la busqueda del Estado del Arte

Fuente	Cadena o palabras clave
WOS	(TS=("credit risk" OR "credit scoring" OR "bank*")) AND (TS=("boosting" OR "logistic regression" OR "alternative data" OR "XAI" OR "explainability" OR "governance" OR "LLM" OR "generative AI"))
SCOPUS	TITLE-ABS-KEY("credit risk" OR "credit scoring" OR "bank*") AND TITLE-ABS-KEY("boosting" OR "logistic regression" OR "alternative data" OR "XAI" OR "explainability" OR "governance" OR "LLM" OR "generative AI")
Semantic Scholar	("credit risk" OR "credit scoring" OR bank OR banking) AND ("boosting" OR "logistic regression" OR "alternative data" OR "XAI" OR "explainability" OR "governance" OR "LLM" OR "generative AI")
ResearchGate	credit risk explainable AI (o credit risk XAI), credit scoring alternative data, banking generative AI, credit risk machine learning

Nota. Fuente: Elaboración propia

Tabla 2.

Criterios de inclusión y exclusión de los papers

Criterios de inclusión	Criterios de exclusión
Artículos científicos publicados en revistas indexadas (ej. Scopus, Web of Science, etc.) o Repositorios confiables de divulgación académica.	Documentos de literatura gris o de divulgación sin renombre ni confiabilidad.

Estudios publicados dentro del periodo comprendido entre 2021 y 2026	Artículos redactados en idiomas distintos al español o inglés.
Investigaciones que apliquen modelos de Machine Learning (ML) a la predicción de riesgo crediticio o casuísticas similares.	Estudios financieros que no utilicen algoritmos de ML (ej. enfoques exclusivamente contables o estadísticos tradicionales).
Artículos que integren o evalúen métodos de explicabilidad e interpretabilidad (XAI) en sus predicciones.	Investigaciones puramente teóricas o simulaciones matemáticas que no validen sus algoritmos con conjuntos de datos reales.

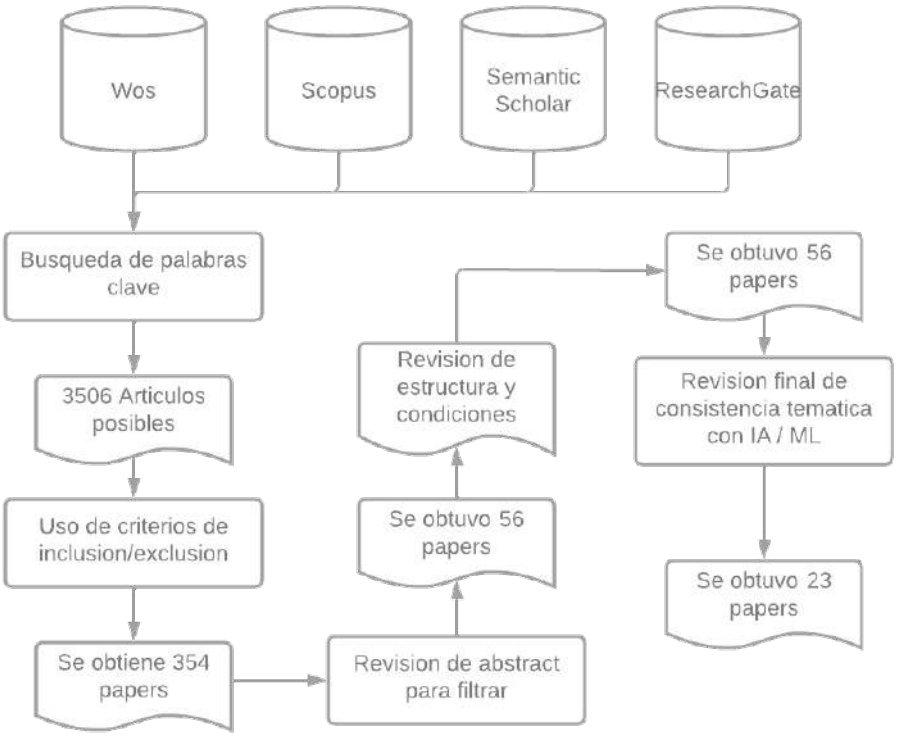
Nota. Fuente: Elaboración propia

3.3 Proceso de recolección

Para la búsqueda de papers se aplicó el criterio de selección utilizado por Wong et al. (2017). Al inicio se buscó identificar los artículos primarios, en el proceso de búsqueda estos artículos fueron sometidos a una selección de acuerdo con los criterios establecidos en la Tabla 2. Para ello, fue necesario hacer una revisión previa del contenido para determinar su pertinencia y finalmente, la mayoría de estos estudios fueron descartados por tratarse de otras áreas como marketing, growth, etc. El flujo de revisión se detalla en la Figura 13.

Figura 13.

Proceso de recoleccion de papers finalistas



Nota. Fuente: Elaboración propia

3.4 Resultados de la recolección

En base a los artículos obtenidos, en la Tabla 3 se puede observar que SCOPUS generó un total de 1869 trabajos para la actual investigación. Por otro lado, los buscadores de WOS, Semantic Scholar y ResearchGate obtuvieron de maneara similar una misma cantidad de artículos elegibles, finalmente los estudios seleccionados en su mayoría provienen de Scopus, ResearcGate y Semantic Scholar.

Tabla 3.

Fuentes por trabajos seleccionados

Fuente	Estudios potencialmente elegibles	Estudios seleccionados
Scopus	1869	9

ResearchGate	654	6
Semantic Scholar	542	5
WoS	441	3
Total	3506	23

Nota. Fuente: Elaboración propia

A continuación, se presenta en la Tabla 4 la lista de 23 papers acompañados del nombre del artículo, referencia, así como también su correspondencia con los ejes tematicos planteados y el cuartil al cual pertenece la revista.

Tabla 4.

Lista de papers seleccionados

ID	Artículo	Referencia	Eje	DB	Cuartil
1	Comparison of Logistic Regression and XGBoost Model Performance in Predicting Credit Scores	Stacyana & Chairunisah (2026)	Eje 1 Algoritmos de boosting	Semantic Scholar	Indexada - Otros
2	Credit Risk Assessment by a Comparison Application of Two Boosting Algorithms	Si et al. (2022)	Eje 1 Algoritmos de boosting	Scopus	Q4
3	Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model	Rao et al. (2023)	Eje 1 Algoritmos de boosting	Scopus	Q1
4	Credit Risk Assessment in P2P Lending Using LightGBM and Particle Swarm Optimization	Dasril et al. (2023)	Eje 1 Algoritmos de boosting	Scopus	Q3
5	Advanced User Credit Risk Prediction Model Using LightGBM, XGBoost and Tabnet with SMOTEENN	Yu et al. (2024)	Eje 1 Balanceo de clases	ResearchGate	Indexada - Otros
6	Some Insights about the Applicability of Logistic Factorisation Machines in Banking	Slabber et al. (2023)	Eje 1 Alternativas interpretables, ensambles y dos etapas	WoS	Q2

7	Two-stage credit scoring using Bayesian approach	Kyeong & Shin (2022)	Eje 1 Alternativas interpretables, ensambles y dos etapas	Scopus	Q1
8	A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from Chinese manufacturing companies	Wang & Zhang (2023)	Eje 1 Alternativas interpretables, ensambles y dos etapas	Scopus	Q1
9	Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring	Zou & Gao (2022)	Eje 1 Alternativas interpretables, ensambles y dos etapas	WoS	Q2
10	On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance	Muñoz-Cancino et al. (2023)	Eje 2 Interacción social y texto	Scopus	Q1
11	Credit default prediction from user-generated text in peer-to-peer lending using deep learning	Kriebel & Stitz (2022)	Eje 2 Interacción social y texto	Scopus	Q1
12	Credit ratings of Chinese online loan platforms based on factor scores and K-means clustering algorithm	Chen et al. (2023)	Eje 2 Calificación no supervisada	Scopus	Q1
13	A SHAP-Based Comparative Analysis of Machine Learning Model Interpretability in Financial Classification Tasks	Chan et al. (2025)	Eje 3 Explicabilidad con SHAP	Scopus	Q4
14	Explainable AI for Credit Scoring with SHAP-Calibrated Ensembles: A Multi-Market Evaluation on Public Lending Data	Japinye & Adedugbe (2025)	Eje 3 Explicabilidad con SHAP	Semantic Scholar	Indexada - Otros
15	SHAP Stability in Credit Risk Management: A Case Study in Credit Card Default Model	Lin & Wang (2025)	Eje 3 Explicabilidad con SHAP	WoS	Q2

16	Fair and Explainable Credit-Scoring under Concept Drift: Adaptive Explanation Frameworks for Evolving Populations	John (2025)	Eje 3 Explicación adaptativa y contextual	ResearchGate	Indexada - Otros
17	Enhancing ML Explainability with Multi-Agent LLMs: A Context-Aware XAI Approach	Miyaji & Corrêa (2025)	Eje 3 Explicación adaptativa y contextual	ResearchGate	Indexada - Otros
18	A Systematic Review of AI-Driven Credit Risk Assessment Models in Commercial Banking (2018–2026)	Sarkar (2026)	Eje 3 Gobernanza y validación	ResearchGate	Indexada - Otros
19	AI and Financial Model Risk Management: Applications, Challenges, Explainability, and Future Directions	Joshi (2025)	Eje 3 Gobernanza y validación	ResearchGate	Indexada - Otros
20	Large Language Models for Automated Financial Code Generation and Documentation in Data Pipelines	Paladugu (2023)	Eje 4 LLM y RAG	Semantic Scholar	Indexada - Otros
21	Agentic Retrieval-Augmented Generation for Financial Document Question Answering	Shu et al. (2026)	Eje 4 LLM y RAG	Semantic Scholar	Indexada - Otros
22	All Leaks Count, Some Count More: Interpretable Temporal Contamination Detection and Mitigation in LLM Backtesting	Zhang et al. (2026)	Eje 4 LLM y RAG	Semantic Scholar	Indexada - Otros
23	Innovations in DataOps for AI in Financial Services: Automation and Efficiency	Boggavarapu (2025)	Eje 4 Infraestructura	ResearchGate	Indexada - Otros

Nota. Fuente: Elaboración propia

3.5 Desarrollo del Estado del Arte

En relación con el primer eje temático planteado (comparación de algoritmos avanzados frente a tradicionales), los estudios que aplican algoritmos de boosting coinciden en que estos

superan a los modelos de referencia. En préstamos automotrices, XGBoost alcanzó una precisión de 96 % y un AUC de 0.99, frente a 88 % y 0.94 de la regresión logística (Stacyana y Chairunisah, 2026). Para otro dataset llamado Lending Club, XGBoost obtuvo un AUC de 0.97, por encima de AdaBoost, regresión logística, análisis discriminante lineal, los árboles de decisión y las máquinas de vectores de soporte (Si et al., 2022).

Esta ventaja puede aumentar si se optimiza los hiperparámetros. Una de las variantes es usar la optimización por enjambre de partículas (PSO). PSO-XGBoost supera a XGBoost estándar, Random Forest y regresión logística en precisión, recall y AUC (Rao et al., 2023). El algoritmo LightGBM con PSO alcanza 98.09 % de precisión y 90.51 % de recall, superando a otras alternativas con casi +3pp adicionales (Dasril et al., 2023). Estos resultados muestran ampliamente que el desempeño depende tanto del algoritmo como de su posterior optimización, tal cual concluyen los 2 anteriores autores.

La revisión de Sarkar (2026) confirma el mismo patrón, los modelos de IA mejoran el AUC entre 4 % y 17 % respecto a la regresión logística. Si bien es cierto las cifras anteriores no permite establecer una jerarquía clara, porque los trabajos usan bases distintas (préstamos automotrices, Lending Club, etc.), algoritmos distintos y técnicas de balanceo distintas. Lo comparable es la dirección del resultado.

Stacyana y Chairunisah (2026) y Si et al. (2022) recomiendan adoptar boosting para discriminar entre buenos y malos solicitantes. Los primeros añaden una salvedad, la regresión logística sigue siendo útil en análisis preliminares donde se exige interpretabilidad lineal.

El desbalanceo aparece como condición transversal. Se aborda con distintas variantes de sobremuestreo, SMOTE (Stacyana y Chairunisah, 2026; Si et al., 2022), Smote-Tomek Link (Rao et al., 2023) y SMOTEENN (Yu et al., 2024). Yu et al. (2024) aplican SMOTEENN con LightGBM, XGBoost y TabNet, y concluyen que el balanceo fortalece la exactitud y la estabilidad, y evita el sobreajuste causado por el desbalance de clases.

Slabber et al. (2023), en cambio, reportan que un bosque aleatorio balanceado presentó sobreajuste claro (Gini de 0.999 en entrenamiento frente a 0.862 en test). Los dos hallazgos no se contradicen, uno se refiere al sobreajuste por desbalance y el otro al del propio modelo. El balanceo, por tanto, no sustituye la evaluación de la estabilidad.

Otro grupo de trabajos no propone reemplazar la regresión logística, sino mejorarla o combinarla. El LFM de segundo orden alcanzó un GINI de 0.829 en validación frente a 0.772 de la logística estándar (Slabber et al., 2023). El modelo logístico bayesiano de dos etapas elevó el KS en 3.42 puntos porcentuales y el AUC en 2.61 (Kyeong y Shin, 2022). Ambos conservan la estructura logística e incorporan interacciones o variables derivadas.

La mejora de Kyeong y Shin (2022) equivale aproximadamente a la mitad de los 5 puntos de AUC que separan a XGBoost de la logística en Stacyana y Chairunisah (2026), aunque con datos distintos. En el mismo estudio, la primera capa también mejoró a Random Forest y a las redes neuronales al colocarse antes de ellos. Wang y Zhang (2023) llevan la misma lógica al crédito empresarial, su ensamble de dos etapas obtuvo el mayor F1 (87.29 %) y la mayor media G (89.47 %) entre los modelos comparados.

Otros resultados muestran que el mejor modelo depende de lo que se priorice. AugBoost-ELM superó a los demás en AUC y precisión en cuatro conjuntos públicos, pero con un menor recall que afectó al F1 y al Brier Score. Por eso Zou y Gao (2022) concluyen que, si se prioriza discriminar entre buenos y malos prestatarios, la regresión logística puede ser suficiente. Si se busca predecir la probabilidad de impago, AugBoost-ELM resulta más adecuado.

Esta conclusión se contrapone parcialmente a la de Stacyana y Chairunisah (2026), que sitúan la logística solo en análisis preliminares. La discrepancia parece provenir del objetivo y de la métrica: ordenar clientes (AUC), detectar incumplidores (recall) o estimar probabilidades (Brier Score). Slabber et al. (2023) prefieren el LFM por su equilibrio entre rendimiento,

estabilidad e interpretabilidad, aunque el GINI en test del bosque aleatorio sea numéricamente mayor.

Pasando al segundo eje temático (los datos alternativos), los estudios coinciden en que estas fuentes aportan valor, pero un valor acotado y dependiente de cuánta información tradicional exista. En un banco latinoamericano, con 97,044 prestatarios y validación cruzada de cinco pliegues, los datos de redes sociales mejoraron la predicción en los primeros meses. Esa ganancia se diluyó cuando el historial de pago ganó relevancia, lo que el análisis SHAP confirmó (Muñoz-Cancino et al., 2023).

En la misma dirección, el texto de los comentarios de los usuarios en redes sociales mejoró en +1 pp el AUC del modelo CNN respecto al uso exclusivo de datos estructurados. Los autores señalan que es especialmente valioso cuando los datos tradicionales son limitados (Kriebel y Stitz, 2022). Ambos trabajos coinciden en que el aporte alternativo crece cuando la información tradicional escasea, sobre todo en públicos desconocidos o llamados no bancarizados. Sin embargo, ambos estudios difieren en el tipo de dato, en el modelo y en el momento en que se mide la ganancia. Los datos alternativos parecen funcionar mejor como complemento en las etapas tempranas de la relación con el cliente que como sustituto de la información de pago.

Por otro lado, Chen et al. (2023) amplían la noción de enfoque no tradicional hacia la calificación sin variable objetivo. Aplicaron análisis factorial y K-means a 130 plataformas de préstamos en línea con 14 indicadores. Los 5 factores extraídos explican 36.16 % de la varianza y el método de la silueta definió 5 grupos. La validación contra datos reales mostró que la proporción de plataformas problemáticas crece a medida que baja la calificación: 0 % en AA, 5.9 % en A, 25 % en BB y 51.4 % en B. A diferencia de los estudios anteriores, aquí se califican entidades y se valida a posteriori.

Ya en relación con el tercer eje temático (terreno de la explicabilidad y la gobernanza), la ventaja predictiva de los modelos complejos ha llevado a varios estudios a acompañarlos con explicaciones SHAP. Todos coinciden en que la complejidad no es incompatible con la transparencia. Chan et al. (2025) compararon cinco modelos con explicaciones SHAP y una partición 80:20. LightGBM ofreció el mejor rendimiento con buena interpretabilidad.

Japinye y Adedugbe (2025) exigen además que la explicación sea auditable. Calibraron XGBoost y LightGBM y controlaron el error falso sobre las explicaciones locales, con calibración casi perfecta (pendiente de 0.98 e intercepto de 0.01) y una correlación temporal de explicaciones superior al 90 % en la mayoría de las auditorías. Lin y Wang (2025) añaden un tercer requisito: la estabilidad frente a la inicialización. Con árboles entrenados bajo distintas semillas, los valores de KS fueron muy cercanos (41 % a 43 %).

Los tres avanzan de forma progresiva: explicar, auditar la explicación y verificar su reproducibilidad. Japinye y Adedugbe (2025) concluyen que los ensambles bien calibrados bajo reglas SHAP ofrecen gobernanza de alto nivel, apta para un escrutinio regulatorio estricto.

Existe otra estrategia, la interpretabilidad por construcción. Conservar la regresión logística cuando la interpretabilidad lineal es un requisito (Stacyana y Chairunisah, 2026), usar modelos concebidos para ser más interpretables que las redes neuronales (Slabber et al., 2023) o recurrir a la logística en etapas (Kyeong y Shin, 2022). Los estudios no comparan directamente ambas estrategias. Queda abierta la pregunta de en qué contextos conviene cada una.

Otro grupo de trabajos desplaza la pregunta hacia la vigencia de la explicación. John (2025) encontró que la interpretación base se degrada con el tiempo, aunque el AUC siga alto. Tres ajustes basados en SHAP restablecen explicaciones coherentes y mitigan variables proxy de atributos protegidos. Este resultado matiza otros hallazgos: Joshi (2025) reporta estabilidad del AUC-ROC (±0.03) y Japinye y Adedugbe (2025) una correlación temporal superior al 90

%. Los hallazgos pueden conciliarse si se considera que miden la estabilidad en condiciones distintas.

Miyaji y Corrêa (2025) abordan otra dimensión: la comunicación. Su sistema multiagente hace que un LLM audite las salidas del modelo clásico y genere justificaciones interactivas en lenguaje natural. Reduce el esfuerzo de auditoría y hace visibles problemas de imparcialidad. Ambos estudios coinciden en que la explicación no es un paso único, sino un proceso que debe mantenerse y comunicarse.

Las dos revisiones coinciden en que el desempeño predictivo no basta. Sarkar (2026) revisó 27 estudios con PRISMA y encontró que los modelos de IA superan a la regresión logística, sobre todo con arquitecturas de conjunto y características transaccionales o temporales. Existe, sin embargo, una brecha creciente entre la innovación y la implementación por calidad de datos deriva, explicabilidad y justicia.

Joshi (2025), con más de 50 fuentes, reporta que entre 50 % y 65 % de las firmas carece de control de versiones y de monitorización en tiempo real. Advierte además que la IA generativa introduce riesgos nuevos, como el origen de los datos. Las recomendaciones son compatibles: marcos de validación estandarizados y explicabilidad alineada a las normativas (Sarkar, 2026), y enfoques híbridos con gobernanza sólida y monitoreo continuo (Joshi, 2025).

Por último, en el cuarto eje temático (terreno de los LLM y la infraestructura), los estudios no se ocupan de estimar el riesgo, sino de automatizar procesos que lo rodean. Coinciden en la necesidad de que los LLM sean verificables. FinCode, ajustado al dominio financiero, generó y documentó código de pipelines y cálculos regulatorios con una certeza regulatoria de 91.2 % frente a 73.2 % de GPT-4 base (Paladugu, 2023).

El marco RAG agentico, con recuperación contrastiva, razonamiento con programas, autoverificación y enrutamiento adaptativo, alcanzó 70.24 % de exactitud con Llama frente a 76.81 % de modelos cerrados más grandes (Shu et al., 2026). Ambos trabajos muestran que la

especialización o la arquitectura mejoran el desempeño, aunque las cifras no son directamente comparables por las distintas tareas y métricas, además ambos estudios lograron implementar el flujo de la IA generativa al corazón de la solución, es decir formo parte de la solución final.

Zhang et al. (2026) aportan un criterio importante para leer estos resultados. Con 2,769 preguntas y la herramienta TimeSPEC con RAG, mostraron que los picos artificiales de rendimiento en backtesting se debían a conocimiento posterior a la fecha de corte. Al eliminar ese sesgo, las estimaciones caen al nivel predictivo real. Esto se suma a la advertencia de Joshi (2025) sobre los riesgos de la IA generativa y a la propuesta de Miyaji y Corrêa (2025) de auditar con un LLM.

Boggavarapu (2025) explora una arquitectura de monitorización automatizada de DataOps capaz de manejar terabytes de inferencia en milisegundos. Reporta que la orquestación adecuada permite procesar más de 100 millones de juicios crediticios en horas. Su enfoque en velocidad y escala es consistente con las cifras de latencia de Paladugu (2023). Sin embargo, esa capacidad técnica contrasta con el dato de Joshi (2025) sobre la ausencia de monitoreo en tiempo real en buena parte de las firmas. La brecha entre lo técnicamente posible y lo implementado sugiere que el límite no es solo tecnológico, sino también de adopción y gobernanza, en línea con lo que señala Sarkar (2026).

3.6 Hallazgos claves

La literatura revisada coincide en cuatro puntos. Los modelos de boosting y los ensambles superan a la regresión logística en capacidad de discriminación (Stacyana y Chairunisah, 2026; Si et al., 2022; Rao et al., 2023; Sarkar, 2026). Los datos alternativos complementan a los tradicionales y aportan más cuando estos escasean (Muñoz-Cancino et al., 2023; Kriebel y Stitz, 2022). La explicabilidad es un requisito para adoptar modelos complejos

(Chan et al., 2025; Japinye y Adedugbe, 2025; Sarkar, 2026; Joshi, 2025). Y la operación exige monitoreo continuo (John, 2025; Joshi, 2025).

Los estudios también se contraponen en cuatro aspectos. El primero es el papel de la regresión logística: útil solo en análisis preliminares para Stacyana y Chairunisah (2026), suficiente según el objetivo para Zou y Gao (2022), o mejorable para Slabber et al. (2023) y Kyeong y Shin (2022). El segundo es la estrategia de interpretabilidad: explicar a posteriori o diseñar modelos que ya sean interpretables. El tercero es el significado de la estabilidad, que puede medirse entre pliegues, entre semillas, entre entrenamiento y prueba, o a lo largo del tiempo. El cuarto es la comparabilidad de las cifras, condicionada por la base de datos, el balanceo y la métrica.

Estas coincidencias y contraposiciones dejan un vacío que la presente investigación busca cubrir. Entre los estudios revisados no se observa un marco integrado que lleve un modelo XGBoost desde su desarrollo hasta su operación explicable y monitoreada, en el contexto de un regulador local como la SBS. El punto de partida está respaldado, pero incluso Stacyana y Chairunisah (2026), que recomiendan XGBoost, reconocen que la regresión logística sigue siendo útil cuando se exige interpretabilidad lineal. Además, las revisiones de Sarkar (2026) y Joshi (2025) documentan una brecha entre la investigación y la implementación asociada a la explicabilidad, la deriva y la falta de monitoreo.

En el plano de la explicabilidad, los estudios con SHAP verifican componentes por separado: la precisión y consistencia de las explicaciones (Chan et al., 2025), su calibración y auditoría (Japinye y Adedugbe, 2025) y su estabilidad frente a semillas aleatorias (Lin y Wang, 2025). Japinye y Adedugbe (2025) son quienes más se acercan a una gobernanza integral, aunque sin abordar el despliegue ni la operación posterior. John (2025) muestra que la interpretación puede degradarse con el tiempo, aunque el AUC siga alto. Joshi (2025) menciona estabilidad del AUC-ROC junto con la ausencia de monitoreo en tiempo real en buena parte de

las firmas. Esto indica que monitorear solo el desempeño predictivo no basta. Falta monitorear también la propia explicación, y las métricas de similitud y de deriva que usa John (2025) resultan un punto de partida útil.

En cuanto a la comunicación, Miyaji y Corrêa (2025) proponen una arquitectura multiagente que mejora las explicaciones en lenguaje natural generadas por un LLM a partir de salidas XAI. Paladugu (2023) y Shu et al. (2026) muestran que la especialización del dominio y el RAG agentico ayudan en la generación de código financiero y en las preguntas sobre documentos financieros. Al mismo tiempo, Zhang et al. (2026) advierten que la evaluación de los LLM puede sobrestimar su desempeño y Joshi (2025) alerta sobre los riesgos de la IA generativa. Por eso una capa LLM requiere ser verificable. En los estudios revisados no se observa que esa capa se articule sobre las explicaciones de un modelo de scoring dentro de un flujo auditable.

Boggavarapu (2025) reporta que la monitorización con indicadores comerciales y métricas operativas puede escalar a más de 100 millones de decisiones, pero sin incorporar la estabilidad de la explicación. Por último, la heterogeneidad de bases, técnicas de balanceo y definiciones de estabilidad limita la comparabilidad entre estudios, y solo uno se ubica explícitamente en un banco latinoamericano (Muñoz-Cancino et al., 2023). Como Sarkar (2026) pide una explicabilidad alineada a las normativas, se justifica un marco reproducible y adaptado a un regulador local.

En consecuencia, la presente investigación propone un modelo XGBoost con un marco completo de desarrollo y explicabilidad, una aplicación web que comunique las explicaciones mediante un LLM robusto, un despliegue escalable en la nube y un monitoreo que incorpore la estabilidad de la explicación. El objetivo es cerrar la brecha entre investigación e implementación. Finalmente, en la Tabla 5 se aprecia un resumen de los hallazgos principales por eje y en la Tabla 6 los vacíos identificados.

Tabla 5.

Resumen de hallazgos claves en papers

Eje	Palabra clave	Hallazgo principal	Autores de referencia
1. Modelos avanzados frente a la regresión logística	Algoritmos de boosting	XGBoost supera a la regresión logística en AUC y precisión (0.99 frente a 0.94 en préstamos automotrices; 0.97 en Lending Club). La optimización con PSO potencia el rendimiento (PSO-XGBoost supera a XGBoost estándar y Random Forest). Las cifras absolutas varían por base, balanceo y métrica.	Stacyana y Chairunisah (2026); Si et al. (2022); Rao et al. (2023); Dasril et al. (2023); Sarkar (2026)
	Manejo del desbalanceo	SMOTE, Smote-Tomek y SMOTEENN son las técnicas usadas. Yu et al. concluyen que SMOTEENN fortalece la exactitud y la estabilidad y evita el sobreajuste por desbalance. Un bosque aleatorio balanceado mostró sobreajuste claro (GINI de 0.999 en entrenamiento frente a 0.862 en test).	Yu et al. (2024); Stacyana y Chairunisah (2026); Si et al. (2022); Rao et al. (2023); Slabber et al. (2023)
	Modelos híbridos y alternativas interpretables	El LFM (GINI de 0.829 frente a 0.772) y el modelo logístico bayesiano de dos etapas (+2.61 puntos de AUC) mejoran a la logística conservando su estructura. Wang y Zhang lograron el mayor F1 en crédito empresarial. AugBoost-ELM logró mayor AUC y precisión, pero menor recall. La elección del modelo depende del objetivo.	Slabber et al. (2023); Kyeong y Shin (2022); Wang y Zhang (2023); Zou y Gao (2022)
2. Datos alternativos	Interacción social y texto	Los datos de redes sociales dan una ganancia en los primeros meses que se diluye cuando el historial de pago gana relevancia. El texto de los usuarios mejora el AUC en un punto cuando los datos tradicionales son limitados.	Muñoz-Cancino et al. (2023); Kriebel y Stitz (2022)
	Calificación no supervisada	El análisis factorial y K-means calificaron 130 plataformas de préstamos. La proporción de plataformas problemáticas sube al bajar la calificación (0 %, 5.9 %, 25 % y 51.4 %).	Chen et al. (2023)

3. Explicabilidad, gobernanza y riesgo de modelos	Explicabilidad con SHAP	SHAP permite explicar modelos complejos con buen rendimiento (LightGBM y XGBoost). La calibración y el control de error falso (bootstrap BCa) dan explicaciones auditables. El modelo es poco sensible a la semilla aleatoria (KS entre 41 % y 43 %).	Chan et al. (2025); Japinye y Adedugbe (2025); Lin y Wang (2025)
	Deriva y explicación adaptativa	La interpretación se degrada con el tiempo, aunque el AUC siga alto. Los ajustes adaptativos con SHAP restablecen explicaciones coherentes. Un LLM puede generar explicaciones en lenguaje natural para audiencias no técnicas.	John (2025); Miyaji y Corrêa (2025)
	Gobernanza y validación	Hay una brecha entre investigación e implementación por calidad de datos, deriva, explicabilidad y justicia. Entre 50 % y 65 % de las firmas carece de control de versiones y de monitorización en tiempo real.	Sarkar (2026); Joshi (2025)
4. LLM e infraestructura	LLM y RAG	FinCode obtuvo 91.2 % de certeza regulatoria frente a 73.2 % de GPT-4 base. El RAG agentico con Llama alcanzó 70.24 % de exactitud frente a 76.81 % de modelos cerrados más grandes. Los picos de rendimiento en backtesting se deben a contaminación temporal.	Paladugu (2023); Shu et al. (2026); Zhang et al. (2026); Joshi (2025)
	Infraestructura	Una orquestación de DataOps con KPI comerciales y métricas operativas permite procesar más de 100 millones de decisiones en horas.	Boggavarapu (2025)

Nota. Fuente: Elaboración propia.

Tabla 6.

Vacios identificados en papers.

Eje	Conclusión	Vacío identificado
1. Modelos avanzados frente a la regresión logística	El boosting y los ensambles superan a la logística en discriminación. Las alternativas interpretables recuperan parte de la brecha. El mejor modelo depende del objetivo (AUC, recall o probabilidad calibrada).	Las cifras no son comparables entre estudios (base, balanceo y métrica). Los resúmenes no precisan si el balanceo se aplicó dentro de cada pliegue. La estabilidad se mide de formas distintas. La interpretabilidad no se usa como criterio de selección junto con el rendimiento.
2. Datos alternativos	Complementan a los datos tradicionales y aportan más cuando estos escasean, sobre todo en la admisión.	Solo un estudio se ubica en un banco latinoamericano. La ganancia es acotada y depende de la etapa del cliente. El estudio no supervisado evalúa plataformas y no prestatarios.
3. Explicabilidad, gobernanza y riesgo de modelos	SHAP hace viable explicar ensambles. La explicación puede degradarse, aunque el AUC se mantenga. La adopción exige explicabilidad y monitoreo.	Cada estudio verifica un aspecto por separado. Ninguno aborda el despliegue ni la operación posterior. No se observa monitoreo de la estabilidad de la explicación en producción. Sigue abierto el debate entre explicar a posteriori e interpretar por diseño. No hay un marco alineado a un regulador local.
4. LLM e infraestructura	La especialización y la arquitectura mejoran el desempeño de los LLM. Su evaluación puede sobrestimarse. La infraestructura ya escala a millones de decisiones.	No se observa una capa LLM articulada sobre las explicaciones de un modelo de scoring en un flujo auditable. Se reporta la infraestructura sin métricas de estabilidad de la explicación. Existe una brecha entre la capacidad técnica y la adopción.
Vacío integrador	Los tres ejes de modelo, explicación y operación se estudian por separado.	No se observa un marco que lleve un modelo XGBoost desde su desarrollo hasta su operación explicable y monitoreo. Se debería incluir una aplicación web con un LLM robusto, despliegue en la nube y monitoreo de la estabilidad de la explicación, en el contexto de un regulador local como la SBS.

Nota. Fuente: Elaboración propia.

CAPÍTULO IV: IMPLEMENTACIÓN

4.1 Desarrollo de la solución

Para el desarrollo de esta investigación se adoptó la metodología CRISP-ML(Q) (Cross-Industry Standard Process for Machine Learning with Quality Assurance). Se trata de una evolución del estándar CRISP-DM, diseñada específicamente para proyectos de machine learning que incorpora aseguramiento de la calidad en cada fase del ciclo de vida (Studer et al., 2021). La principal diferencia con su predecesor es el énfasis en la gestión de riesgos y el monitoreo continuo del modelo en producción. Este enfoque es crítico para garantizar la fiabilidad y el rendimiento sostenido del sistema en el entorno productivo. CRISP-DM, en cambio, no aborda esta necesidad de supervisión operativa. La Figura 14 muestra los pasos de esta metodología aplicada al presente estudio.

Figura 14.

Las 6 Fases de Metodologia CRISP-ML(Q)



Nota. Fuente: Elaboración propia

A continuación, en la Tabla 7, se detalla cada fase de la metodología, su concepto y su relación con los objetivos específicos de la investigación.

Tabla 7.

Fases de Metodologia CRISP-ML(Q) relacionadas a los objetivos especificos

Fase CRISP-ML(Q)	Descripción	OE	Fase de Desarrollo
1. Entendimiento del Negocio y Datos	Comprensión del problema, definición de objetivos y análisis exploratorio de los datos disponibles.	OE1	Construcción de dataset inicial con variables de carácter crediticio social y comportamental
2. Preparación y Feature Engineering	Limpieza, transformación y creación de variables predictoras a partir de los datos crudos.	OE2	Preprocesamiento, limpieza y análisis exploratorio del dataset inicial
3. Desarrollo y Evaluación	Entrenamiento, ajuste de hiperparámetros y validación del rendimiento del modelo.	OE3	Evaluación, entrenamiento y selección del mejor modelo predictivo de impago para Préstamos Personales
4. Despliegue e Integración	Puesta en producción del modelo, integración con sistemas existentes y exposición mediante APIs.	OE4	Diseño e implementación del sistema de seguimiento de performance, estabilidad e interpretabilidad del modelo
5. Monitoreo y Mantenimiento	Supervisión continua del rendimiento, detección de deriva (drift) y actualización del modelo.		

Nota. Fuente: Elaboración propia

4.1.1 Construcción del dataset inicial con variables de carácter crediticio, social y comportamental

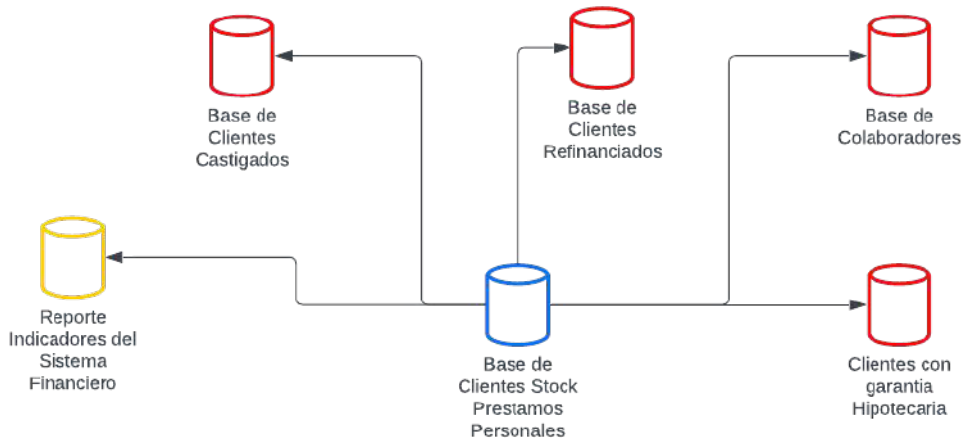
4.1.1.1 Recolección de la Poblacion Objetivo y Definición del Horizonte Temporal

Antes de recolectar los datos de la población objetivo (PO), se construyeron los filtros necesarios para excluir a personas con marcas indefinidas que no aportan a la discriminación del modelo. Albónico (2021) indica la importancia de contar con poblaciones similares al momento de modelar, ya que el comportamiento errático puede causar fluctuaciones

importantes sobre el output del modelo. Durante el proceso de recolección se siguió la estructura de la Figura 15.

Figura 15.

Estructura para recolección de fuentes

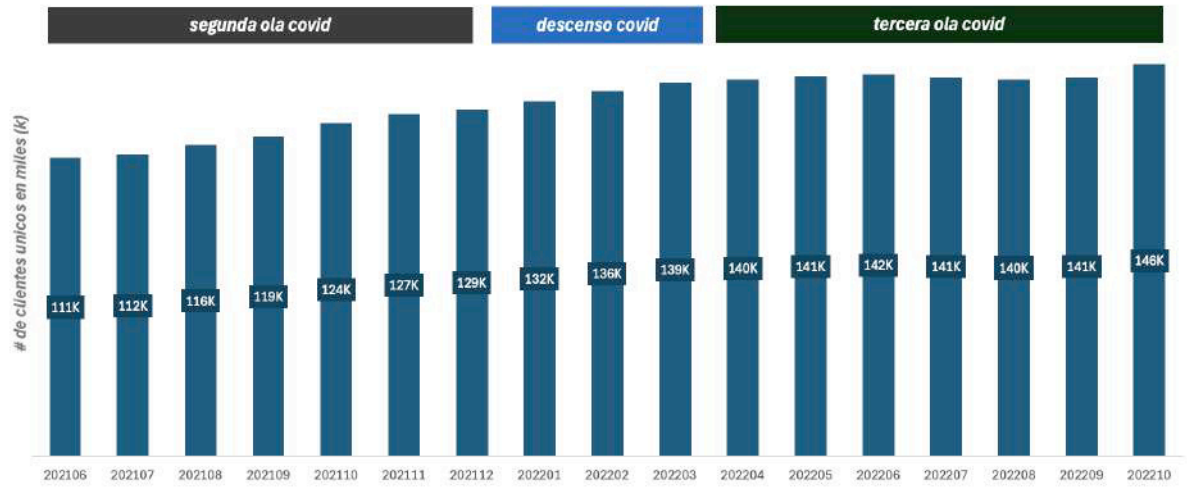


Nota. Fuente: Elaboración propia

Las bases de datos en color rojo (castigados, refinanciados, colaboradores, garantía hipotecaria) sirven para realizar un filtro en la población objetivo inicial. La base de datos en color amarillo (Indicador SSFF) se utiliza tanto para construir filtros como para la construcción de variables en los pasos siguientes. El periodo escogido para la recolección de la población objetivo abarca desde 202106 hasta 202210. En este rango se busca capturar el comportamiento de los clientes en periodos de pandemia y postpandemia de COVID-19 en el Perú. El objetivo es que el modelo sea más robusto ante situaciones similares. La población inicial recolectada se refleja en la Figura 16.

Figura 16.

Numero de clientes con prestamos presonaes durante 2021-2022



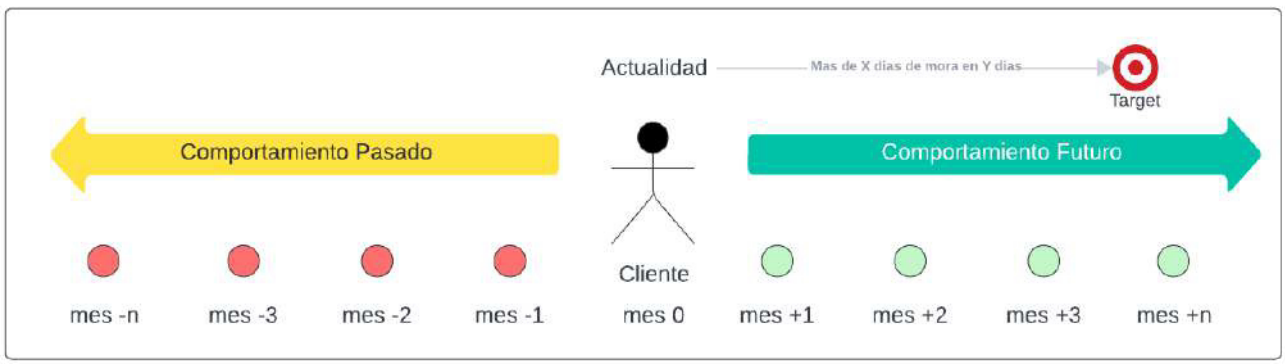
Nota. Fuente: Elaboración propia

4.1.1.2 Construcción y Definición de la Variable Objeto (Target) mediante Analisis de Roll Rates

Uno de los puntos más importantes al construir el dataset de cualquier modelo es establecer el target (variable objetivo a predecir). Para este modelo, se define como target la probabilidad de incumplimiento (probabilidad de default) en la cartera de préstamos personales. El correcto planteamiento del target garantiza un buen inicio para que el modelo funcione bien. La Figura 17 explica de manera gráfica el target del modelo.

Figura 17.

Diagrama de construcción y comportamiento del target



Nota. Fuente: Elaboración propia

En la figura anterior se explica el objetivo del target, predecir en el futuro qué clientes se van a atrasar en el pago de su producto de préstamos personales. Para calcular los valores de X e Y se aplicó un análisis de Roll Rates. El análisis de Roll Rates permite detectar que punto es óptimo para que un cliente ya no se recupere y por lo tanto se marce como default. Se realizó un análisis de Roll Rates para los primeros meses de la PO (enero - mayo 2021), comparando los días de mora del mes m0 respecto al m+1, tal como se observa en la Tabla 8. La diagonal superior indica la tasa de deterioro máxima en una ventana de 12 meses, mientras que la diagonal inferior es la tasa de recuperación. En este análisis se aprecia que el rango de Mora 61-75 y 76-90 obtiene la mayor tasa de deterioro.

Tabla 8.

Analisis de Roll Rates – Matriz de transición entre inicio y fin

Final Inicio	000-000	001-007	008-015	016-030	031-045	046-060	061-075	076-090	091-105	106-120	> 121
000-000	87.6%	1.0%	2.2%	2.1%	0.7%	1.1%	0.4%	0.7%	0.4%	1.4%	2.6%
001-007	13.6%	61.7%	1.9%	0.4%	13.6%	0.0%	3.0%	0.0%	3.0%	0.4%	2.3%
008-015	16.3%	0.1%	20.6%	0.3%	13.9%	0.5%	6.3%	0.2%	8.4%	0.6%	32.9%
016-030	17.1%	0.1%	0.2%	13.9%	0.1%	12.7%	0.1%	7.1%	0.0%	15.5%	33.2%
031-045	9.7%	2.1%	3.5%	0.0%	9.3%	0.2%	11.1%	0.0%	5.4%	7.8%	51.0%
046-060	11.1%	0.0%	0.0%	2.9%	0.2%	6.6%	0.0%	11.6%	0.0%	4.7%	63.1%
061-075	8.7%	0.7%	1.8%	0.0%	2.2%	0.2%	8.2%	0.0%	16.3%	3.6%	58.4%
076-090	12.7%	0.0%	0.0%	1.4%	0.0%	1.4%	0.0%	5.1%	0.2%	18.1%	61.3%

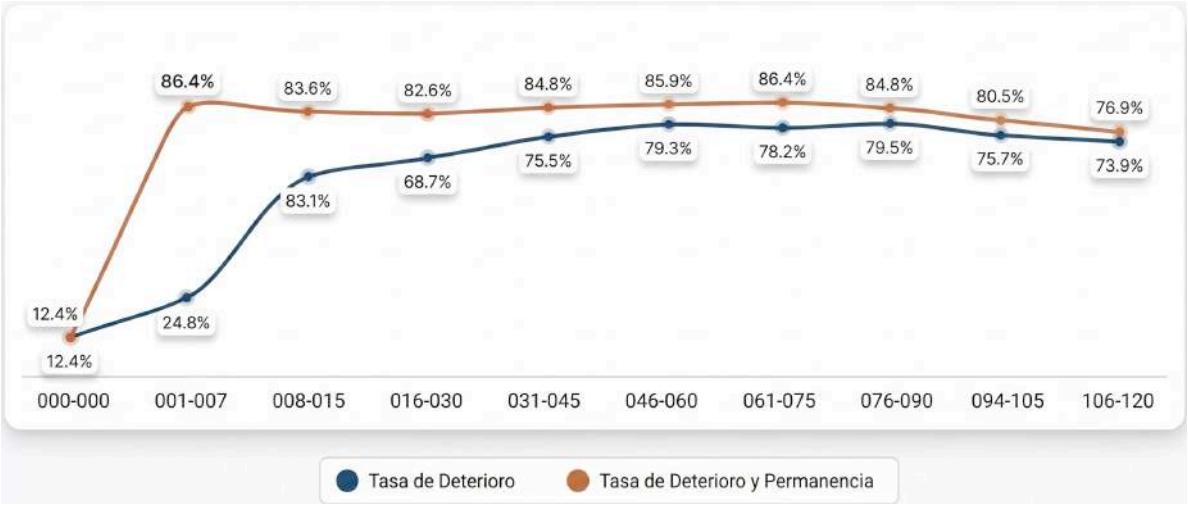
091-105	14.9%	0.3%	0.8%	0.0%	1.1%	0.0%	2.4%	0.0%	4.8%	0.8%	74.9%
106-120	21.5%	0.0%	0.0%	0.4%	0.0%	0.6%	0.0%	0.7%	0.0%	3.0%	73.9%
> 121	88.0%	0.4%	0.0%	0.0%	0.1%	0.0%	0.0%	0.1%	0.4%	0.4%	10.5%
Total	84.1%	1.1%	2.3%	2.1%	0.9%	1.2%	0.5%	0.8%	0.5%	1.6%	4.9%

Nota. Fuente: Elaboración propia

A partir de la tabla anterior se concluye que el mejor target es aquel que demuestre una gran probabilidad de que el cliente no pueda recuperarse, es decir, que no llegue a pagar su préstamo. Como lo representa la Figura 18, este valor corresponde al límite de 75 días de mora, que es el intervalo en donde se estabiliza la tasa de deterioro. Por lo tanto, se establece como target: más de 75 días de mora en una ventana de 12 meses futuros. A estos clientes se les conocerá como "Malos". Al resto de clientes que tengan menos de 75 días de mora se les conocerá como "Buenos".

Figura 18.

Curva de Tasa de Deterioro – Rol Rates



Nota. Fuente: Elaboración propia

Sin embargo, existe un segmento de clientes que tienen entre 7 y 75 días de mora. Estos tienen una probabilidad media de recuperarse. A estos clientes los nombraremos como "vaivén", porque su comportamiento es confuso. Los clientes "buenos netos" serán los de 0 a 7 días de mora.

4.1.1.3 Aplicación de Filtros de Inclusión y Exclusión a la Población Objetivo

Durante este proceso se realizó una segmentación de los clientes para obtener la población más limpia posible. Para ello, se desarrollaron una serie de filtros, tal como lo muestra la Tabla 9.

Tabla 9.

Filtros Aplicados a la Poblacion Objetivo – Criterios de negocio y calidad

Nombre del Filtro	Descripción	Tipo de Filtro	¿Por qué filtrar?
Clientes vaivén	Clientes sin comportamiento definido, impredecibles	Exclusión	Comportamiento confuso; misma probabilidad de recuperarse que de empeorar. No suma al modelo.
Malos en el Parque Financiero	Clientes con etiqueta de malo en todo el sistema financiero	Exclusión	Ya son malos para otros bancos; no tiene sentido predecir su morosidad.
Colaborador	Clientes que también son colaboradores de la institución	Exclusión	Comportamiento peculiar por conocer el negocio de la banca.
Malos de raíz	Clientes que ya son malos desde el mes de observación	Exclusión	Ya son malos en el mes de observación; no tiene sentido predecir su futuro.

Clientes con Garantía Hipotecaria	Clientes con hipoteca como aval	Exclusión	Comportamiento distinto al tener garantía hipotecaria.
Clientes finiquitados	Clientes sin deuda pendiente en el mes de observación	Exclusión	Comportamiento diferente al no tener deuda.
Sin Comportamiento	Clientes sin comportamiento en más de 9 meses a futuro	Exclusión	No se puede saber su comportamiento por falta de periodos completos.
Cliente Bueno	Clientes con menos de 75 días de mora en los siguientes 12 meses	Inclusión	Futuro comportamiento favorable.
Cliente Malo	Clientes con 75+ días de mora en los siguientes 12 meses	Inclusión	Partieron como buenos, pero se transformaron en morosos o fueron castigados.

Nota. Fuente: Elaboración propia

Una vez definidos los filtros, se procedió a la exclusión de clientes bajo las diferentes categorías. En la Tabla 10 se aprecia el resultado de ir excluyendo clientes filtro tras filtro (filtro cascado). Al final, la P.O. limpia (Buenos y Malos) quedó en 1,735,340 clientes.

Tabla 10.

Filtro Cascada de la Poblacion Objetivo

Filtro Cascada	# a Retirar	# Restante
Base Inicial	-	2,235,642
Clientes Vaivén	171,220	2,064,422
Malos Parque Financiero	79,903	1,984,519
Colaboradores	37,071	1,947,448
Malos de Raíz	11,171	1,936,277
Garantía Hipotecaria	4,998	1,931,279
Clientes Finiquitados	1,768	1,929,511

Sin Comportamiento	194,171	1,735,340
PO Final	-	1,735,340
Buenos	-	1,530,737
Malos	-	204,603

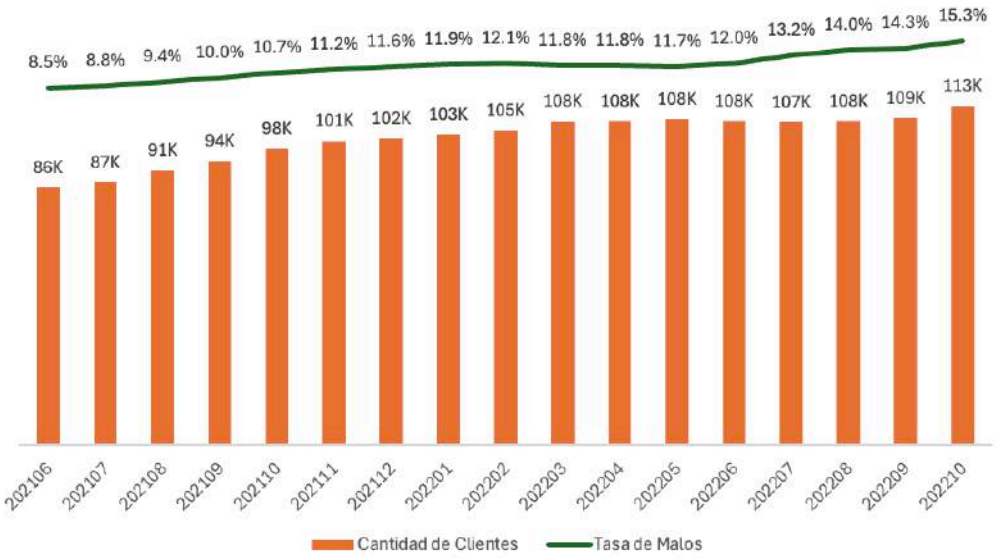
Nota. Fuente: Elaboración propia

4.1.1.4 Evaluación de la Estabilidad de la Tasa de Malos

Antes de realizar la construcción de variables, se validó la estabilidad de la tasa de malos (número de clientes con etiqueta Malo / número total de clientes). En la Figura 19 se observa que el número de clientes ha ido aumentando conforme avanzan los meses. La tasa de malos también tuvo un incremento, explicado por el aumento del stock de clientes. Como no existe un cambio drástico en la tasa de malos, se puede considerar que esta población objetivo final es estable para modelar. Además, recoge el comportamiento de crisis de la pandemia de COVID-19 en el Perú.

Figura 19.

Evolutivo de tasa de malos por mes



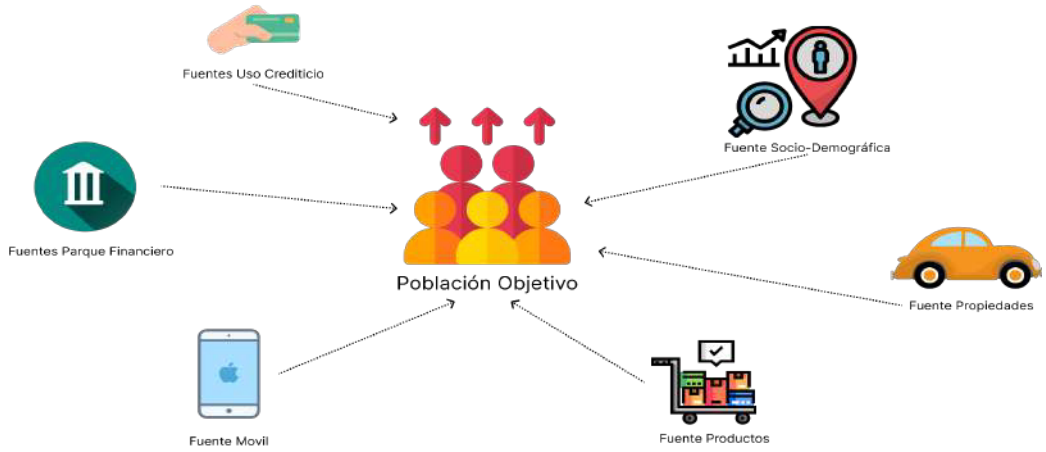
Nota. Fuente: Elaboración propia

4.1.1.5 Recolección y Distribución de Variables por Fuentes de Información

Durante el desarrollo de esta sección se procedió a construir querys a partir de las diferentes fuentes de información, logrando finalmente un total de 582 variables. En la Figura 20 se visualiza el flujo general de construcción y las fuentes de información consultadas.

Figura 20.

Fuentes Consultadas para la creación del dataset



Nota. Fuente: Elaboración propia

La fuente de Uso crediticio tiene un 52% de representación. Contiene variables como % de uso de Tarjeta de crédito, número de meses en los que el cliente hizo un decremento de su saldo, entre otras. La fuente Parque Financiero es similar a Uso crediticio, pero incluye al resto de instituciones financieras. Las fuentes Productos, Sociodemográficas, Móviles y Propiedades recolectan información del cliente en cada categoría: nivel de educación, tipo de móvil, tenencia de propiedades, tenencia de productos, etc. La Tabla 11 presenta de manera resumida la distribución y número de variables por fuente

Tabla 11.

Peso de Variables por fuente de información

Nombre de la Fuente	Número de Variables	% Representación
Uso crediticio	305	52%
Parque Financiero	96	16%
Productos	84	14%
Sociodemográfico	42	7%
Móvil	27	5%
Propiedad	28	5%
Total Fuentes	582	100%

Nota. Fuente: Elaboración propia

4.1.2 Preprocesamiento, limpieza y análisis exploratorio del dataset inicial

4.1.2.1 Filtros Generales de Reducción de Dimensionalidad (Missings, Gini Mínimo y Correlación)

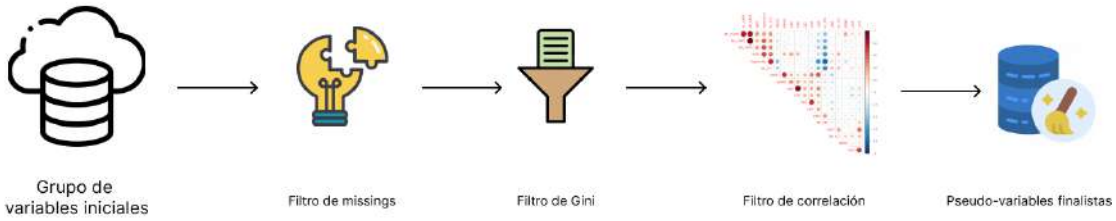
Se aplicaron tres filtros estrictos a las 582 variables iniciales:

- Número de Missings: se retiraron variables con más de 95 % de valores faltantes.
- Gini mínimo: se eliminaron variables con un índice de Gini inferior a 5 %.
- Grado de correlación: se descartaron variables con más de 60 % de correlación entre sí.

El diagrama de la secuencia de pasos se aprecia en la Figura 21, donde se detalla el orden de aplicación de los filtros.

Figura 21.

Flujo de fitros para dataset inicial de variables



Nota. Fuente: Elaboración propia

Por razones prácticas, el dataset inicial de 582 variables se dividió en cuatro grupos Grupo b (Uso crediticio), Grupo b2 (Uso crediticio + Parque financiero), Grupo cd (Sociodemográficos) y Grupo ed (Producto, Móvil y Propiedad). Durante este proceso solo se analizaron variables numéricas (continuas y categóricas-ordinales). Las variables object se tratarán en apartados posteriores. La cantidad de variables que superaron cada filtro se detalla en la Tabla 12.

Tabla 12.

Set de variables disgregado por grupo y filtro aplicado

Grupo	Total Inicial	Filtro Missing	Filtro Gini	Filtro Correlación	Object	Total Final
b	154	40	38	59	0	17
b2	125	47	20	42	0	16
cd	199	24	36	131	0	8
ed	104	35	45	16	2	6
						47

Nota. Fuente: Elaboración propia

Después de aplicar los filtros por grupo, se obtuvieron 17 variables para el grupo b, 16 para b2, 8 para cd y 6 para ed. Luego se unieron todas las variables y se volvió a ejecutar el proceso completo, porque podían existir correlaciones entre variables de distintos grupos. De

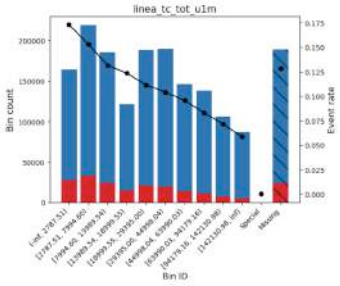
las 47 variables, se retiraron 4, quedando finalmente 43 variables. Las variables básicas como edad y sexo tienen una fuerte relación con el ingreso del cliente y, a su vez, con la probabilidad de default. Por ello, a las 43 variables finalistas se agregaron 8 campos adicionales (edad, sexo, distrito de residencia, número de hijos, etc.). El dataset quedó con 51 variables tentativas.

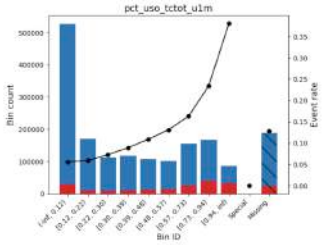
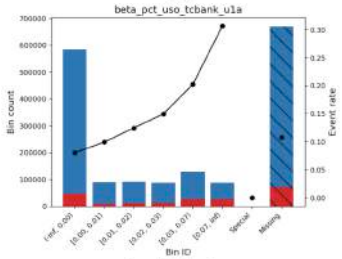
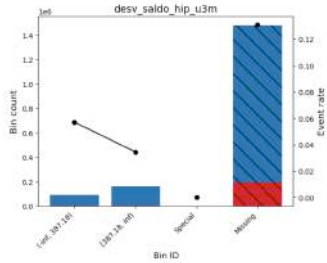
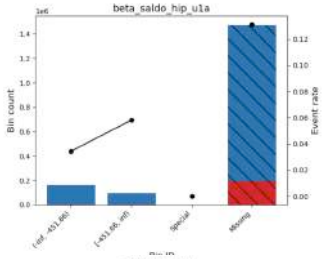
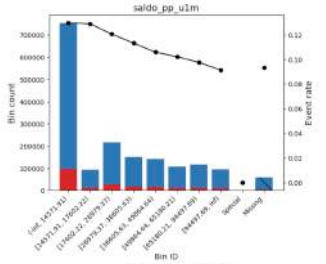
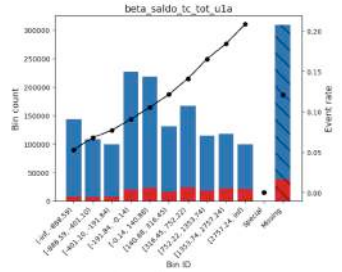
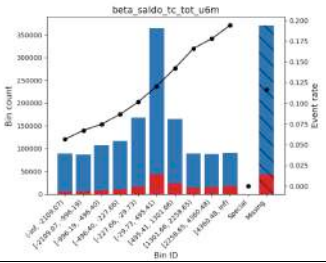
4.1.2.2 Análisis Bivariado y Transformación de Variables mediante Peso de la Evidencia (WOE)

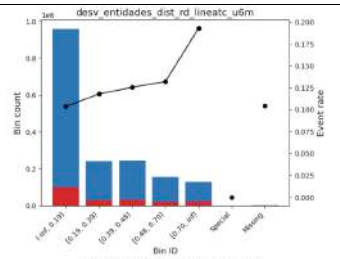
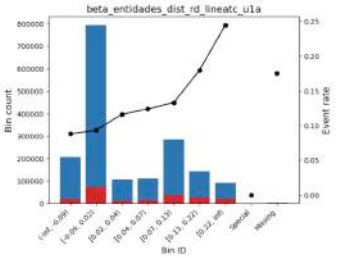
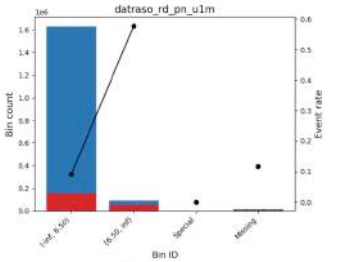
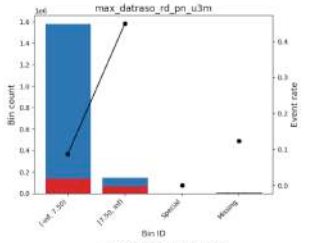
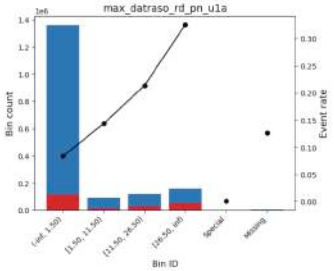
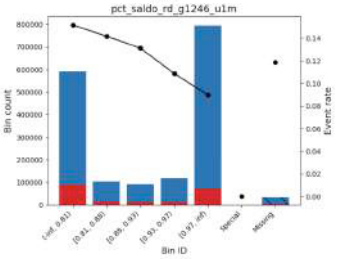
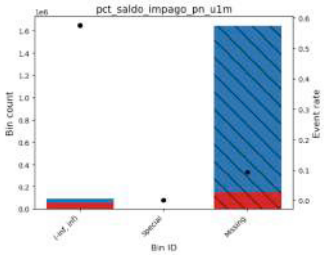
Con estas 51 variables se realizaron dos pasos finales. Primero, se ejecutó un análisis bivariado para cada variable en función del target, con imputaciones cuando fue necesario. Segundo, se entrenó un toy model con las variables imputadas, eliminando aquellas con importancia insignificante. Se utilizó la librería optbinning para realizar cortes óptimos que permiten visualizar el sentido de cada variable en función del target. El objeto OptimalBinning se personalizó para que los cortes fueran monótonos (ascendentes o descendentes), con una representatividad mínima del 5 % y máxima del 95 %, asegurando bins representativos. En la Tabla 13 se muestran los resultados del proceso y su interpretación para las 49 variables continuas del dataset.

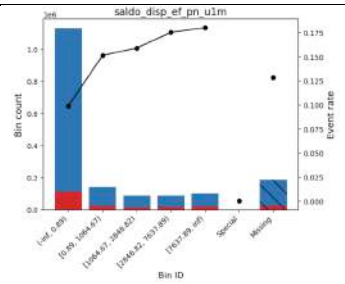
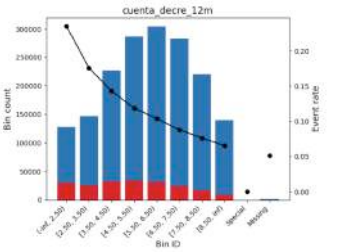
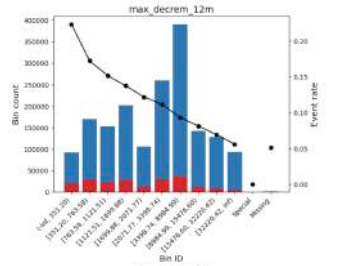
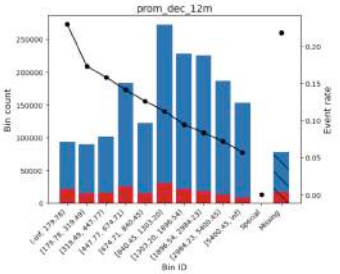
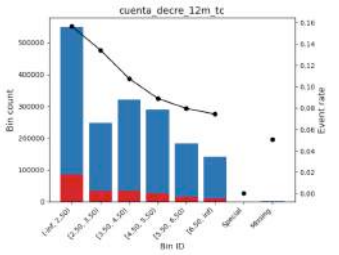
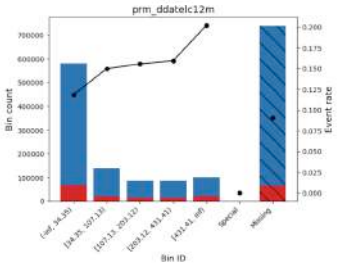
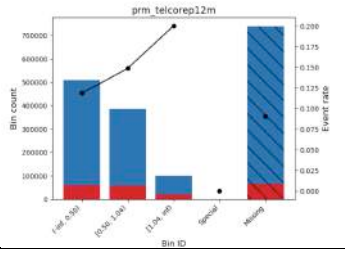
Tabla 13.

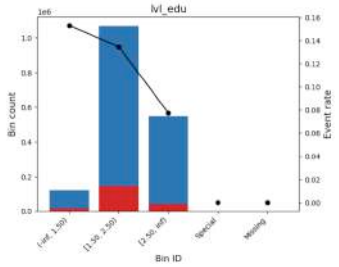
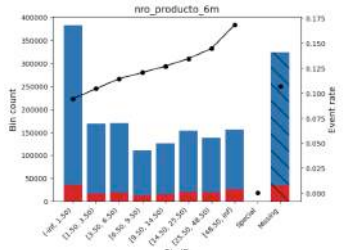
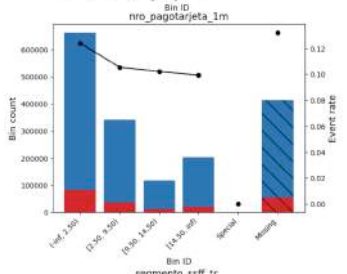
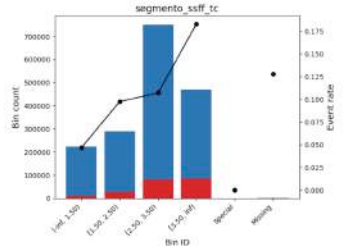
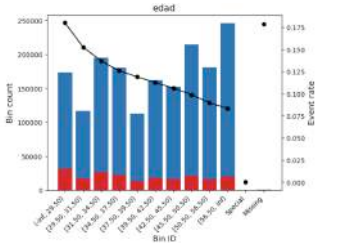
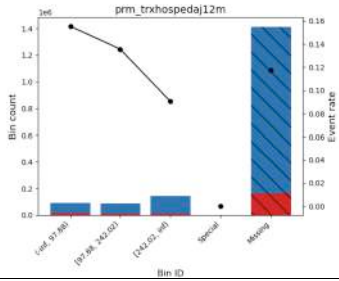
Diccionario de Variables Filtradas y grafico Bivariado

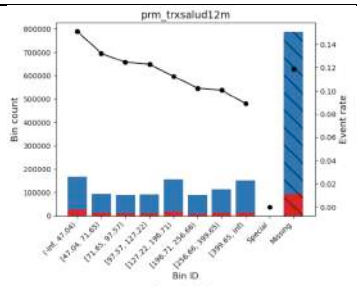
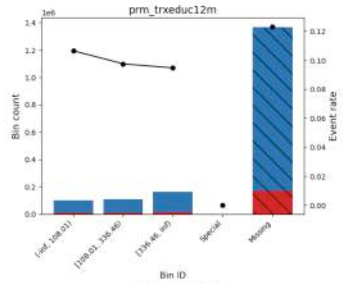
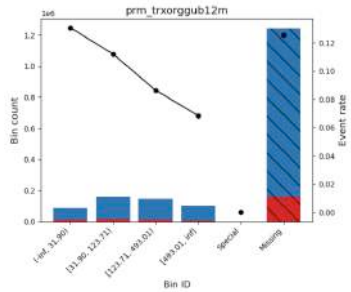
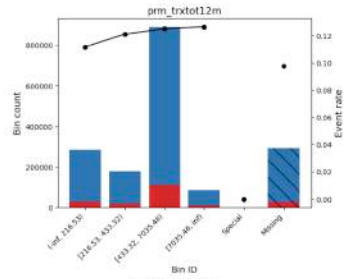
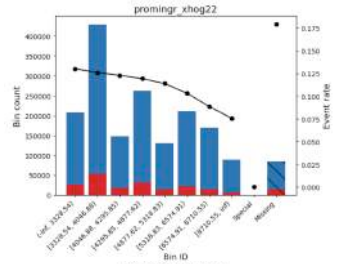
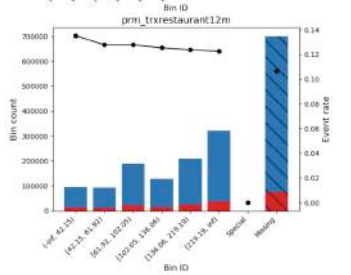
Gráfico Bivariado	Variable	Interpretación
	Línea TC Total del último mes	Variable relación inversa con el target. A mayor línea de TC menor riesgo asociado

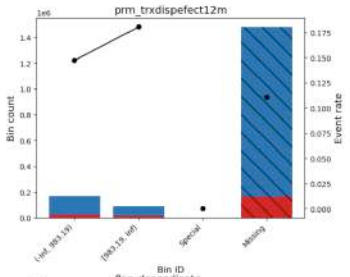
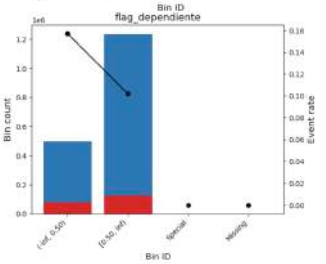
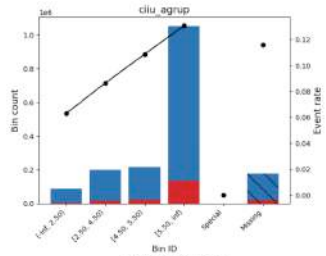
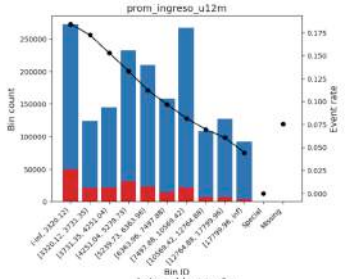
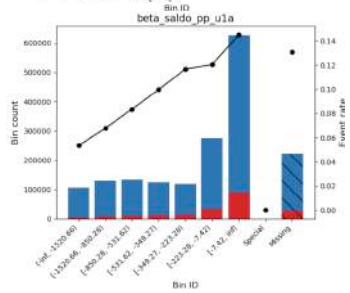
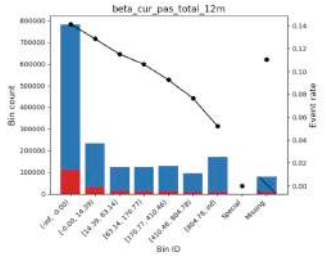
	Porcentaje de uso Tc total último mes	Variable relación directa con el target. A mayor % de uso línea de TC mayor riesgo asociado.
	Beta porcentaje de uso tc último año	Variable relación directa con el target. Cuando la tendencia del pct de uso aumenta existe un mayor riesgo asociado.
	Desviación saldo hipotecario últimos 3 meses	Variable relación inversa con el target. A mayor variación (disminución de la deuda) menor riesgo asociado
	Beta saldo hipotecario último año	Variable relación directa con el target. A menor tendencia a disminuir (negativo) mayor riesgo asociado.
	Saldo préstamos personales último mes	Variable relación inversa con el target. A mayor saldo préstamos personales menor riesgo asociado.
	Beta saldo Tarjeta de crédito total último año	Variable relación directa con el target. A mayor tendencia de aumentar la deuda, mayor riesgo asociado.
	Beta saldo Tarjeta de crédito total últimos 6 meses.	Variable relación directa con el target. A mayor tendencia de aumentar la deuda, mayor riesgo asociado.

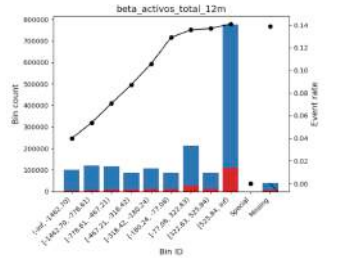
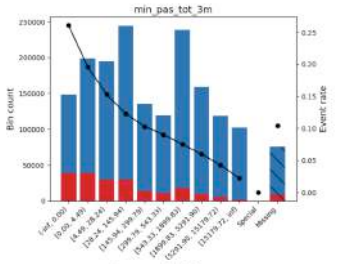
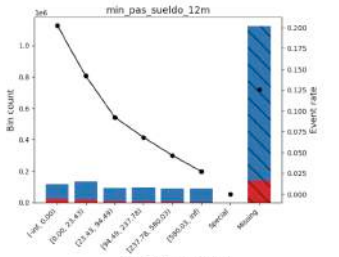
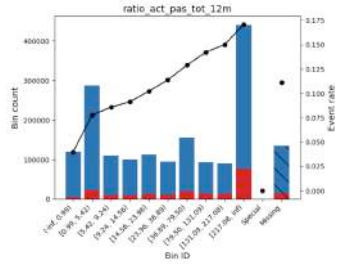
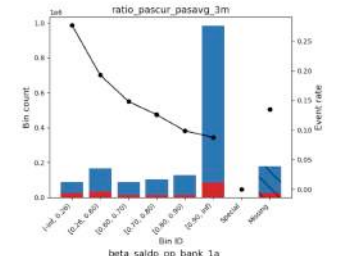
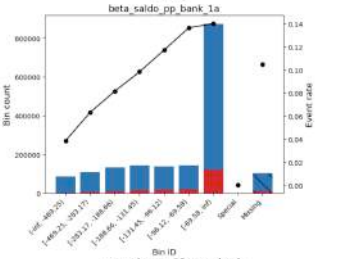
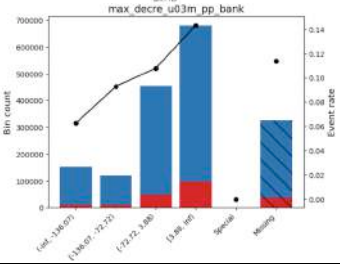
	Desviación % cantidad de entidades con deuda en los últimos 6 meses	Variable relación directa con el target. A mayor desviación con la cantidad de entidad con las que tengo deuda mayor riesgo asociado.
	Beta entidades con deuda tarjeta de crédito último año	Variable relación directa con el target. A mayor tendencia a aumentar la cantidad de entidades donde tengo deuda de tc mayor riesgo asociado.
	Días de atraso riesgo directo persona natural	Variable relación directa con el target. A mayores días de atraso en el último mes mayor riesgo asociado.
	Máximo días de atraso riesgo directo persona natural últimos 3 meses	Variable relación directa con el target. A mayor máximo días de atraso en los últimos 3 meses mayor riesgo asociado.
	Máximo días de atraso riesgo directo persona natural último año	Variable relación directa con el target. A mayor máximo días de atraso en el último año mayor riesgo asociado.
	Porcentaje de saldo riesgo directo en principales entidades	Variable relación inversa con el target. A mayor porcentaje de deuda en las principales entidades menor riesgo asociado.
	Porcentaje de saldo impago en el último mes	Variable no presenta una tendencia clara con el target. (Se eliminará del dataset)

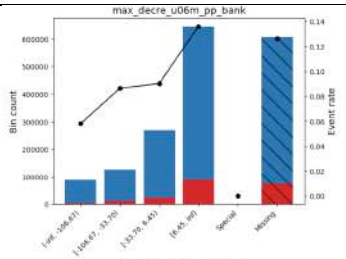
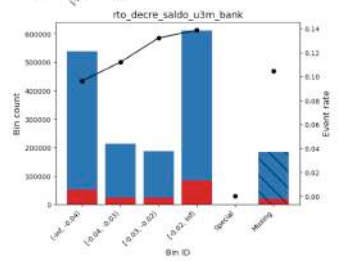
	Disposición de efectivo tarjeta de crédito en el último mes	Variable relación directa con el target. A mayor disposición de efectivo mayor riesgo asociado.
	Número de decrementos en el último año	Variable relación inversa con el target. A mayor número de decrementos (pagos de deuda) menor riesgo asociado.
	Máximo decremento en el último año	Variable relación inversa con el target. A mayor máximo del monto pagado en el último año, menor riesgo asociado.
	Promedio decremento último año.	Variable relación inversa con el target. A mayor promedio del monto pagado en el último año, menor riesgo asociado.
	Número de decrementos en Tarjeta de crédito en el último año.	Variable relación inversa con el target. A mayor número de decrementos de tarjeta de crédito en el último año, menor riesgo asociado.
	Promedio de deuda telecomunicaciones último año	Variable relación directa con el target. A mayor deuda reportada en telecomunicaciones mayor riesgo asociado.
	Promedio número de empresas que reportan deuda en telecomunicaciones último año	Variable relación directa con el target. A mayor promedio de número de empresas que reportan deuda en telecomunicaciones mayor riesgo asociado.

	Variable de nivel educativo: 1 = Iletrados, Primaria, Secundaria incompleta, Nulos 2 = Secundaria completa, Educación técnica incompleta 3 = Educación técnica completa, educación superior	Variable relación inversa con el target. A mayor nivel educativo menor riesgo asociado.
	Números de intenciones de acceder al producto en los últimos 6 meses.	Variable relación directa con el target. A mayores intenciones de acceder a productos, mayor riesgo asociado.
	Número de pagos de tarjetas en el último mes	Variable relación inversa con el target. A mayor cantidad de pagos de tarjeta en el último mes, menor riesgo asociado.
	Segmento Sistema financiero TC, en función del riesgo	Variable relación directa con el target. A mayor segmento de tc, mayor riesgo asociado.
	Edad del cliente	Variable relación inversa con el target. A mayor edad, menor riesgo asociado.
	Promedio transacciones hospedajes último año	Variable relación inversa con el target. A mayor gasto en hospedajes, menor riesgo asociado.

	Promedio transacciones salud último año	Variable relación inversa con el target. A mayor gasto en salud, menor riesgo asociado
	Promedio transacciones educación último año	Variable relación inversa con el target. A mayor gasto en educación, menor riesgo asociado
	Promedio transacciones organizaciones gubernamentales el último año.	Variable relación inversa con el target. A mayor gasto en org gubernamentales, menor riesgo asociado
	Promedio transacciones totales en el último año.	Variable relación inversa con el target. A mayores transacciones totales, menor riesgo asociado
	Promedio gasto mensual del hogar entre año 2016 y 2022	Variable relación inversa con el target. A mayor gasto mensual en el hogar, menor riesgo asociado.
	Promedio transacciones en restaurantes en el último año.	Variable relación inversa con el target. A mayores transacciones en restaurantes, menor riesgo asociado.

	Promedio transacciones y disposición de efectivo en el último año.	Variable relación directa con el target. A mayores transacciones de DE, mayor riesgo asociado.
	Flag dependiente 0 = Independiente 1 = Dependiente	Variable relación inversa con el target. Si es dependiente menor riesgo asociado.
	Agrupación actividad económica	Variable relación directa con el target. Actividad económica con mayor riesgo registrado en sunat mayor riesgo financiero asociado.
	Promedio ingreso últimos año	Variable relación inversa con el target. A mayor ingreso, menor riesgo asociado.
	Beta saldo préstamos personales último año	Variable relación directa con el target. A menor tendencia a disminuir la deuda mayor probabilidad de riesgo asociado.
	Beta Pasivo a fin de mes últimos 12 meses	Variable relación inversa con el target. A mayor tendencia a aumentar los pasivos a fin de mes menor riesgo asociado.

	Beta activos total últimos 12 meses	Variable relación directa con el target. A mayor tendencia a aumentar la cantidad de saldo de activos, mayor riesgo asociado.
	Mínimo pasivo total últimos 3 meses	Variable relación inversa con el target. A mayor mínimo de pasivo total, menor riesgo asociado.
	Mínimos pasivos en cuenta sueldo último año	Variable relación inversa con el target. A mayor mínimo de pasivo en cuenta sueldo, menor riesgo asociado.
	Ratio entre activos y pasivos de los últimos 12 meses	Variable relación directa con el target. Cuando el valor de los activos es mucho más grande que los pasivos existen un mayor riesgo asociado.
	Ratio pasivos fin de mes entre pasivos mensuales últimos 3 meses.	Variable relación inversa con el target. Cuando el valor de los activos a fin de mes es similar a los mensuales existe un menor riesgo.
	Beta saldo préstamos personales último año	Variable relación directa con el target. La tendencia a disminuir el pago el saldo aumenta, existe un mayor riesgo.
	Máximo decremento (negativo) Préstamos personales en los últimos 3 meses	Variable relación directa con el target. A mayor valor de la variable en negativo, existe un mayor riesgo.

	Máximo decremento (negativo) Préstamos personales en los últimos 6 meses	Variable relación directa con el target. A mayor valor de la variable en negativo, existe un mayor riesgo.
	Ratio decremento entre saldo últimos 3 meses	Variable relación directa con el target. Cuando el decremento en negativo sea menor en comparación al saldo, existe un mayor riesgo.

Nota. Fuente: Elaboración propia

Del análisis anterior, la variable pct_saldo_impago_pn_u1m no mostró una tendencia clara con el target, por lo que fue eliminada del dataset. La librería optbinning también permitió transformar variables categóricas bajo el enfoque WOE. En el dataset hay dos variables categóricas que no están en la tabla anterior: Procedencia profesional y Segmentación RFM. En las siguientes Tablas 14 y 15, y las Figuras 22 y 23 se muestra la transformación y sus respectivos gráficos bivariados. Interpretando esos gráficos se entiende que a mayor valor de WOE (Bin) menor probabilidad de riesgo, de esta manera se redujo la dimensionalidad de la variable a grupos riesgosos.

Tabla 14.

Transformación WOE para la variable Procedencia Profesional

Bin	Count	Count (%)	Event rate	WoE	IV
[PRIVADA]	153093	0.10	0.053	0.82	0.049
[NACIONAL, TECNICO]	169782	0.11	0.061	0.66	0.038
Special	0	0	0	0	0
Missing	1191224	0.79	0.12	-0.142	0.016

Total	1514099	1	0.113	0.10

Nota. Fuente: Elaboración propia

Tabla 15.

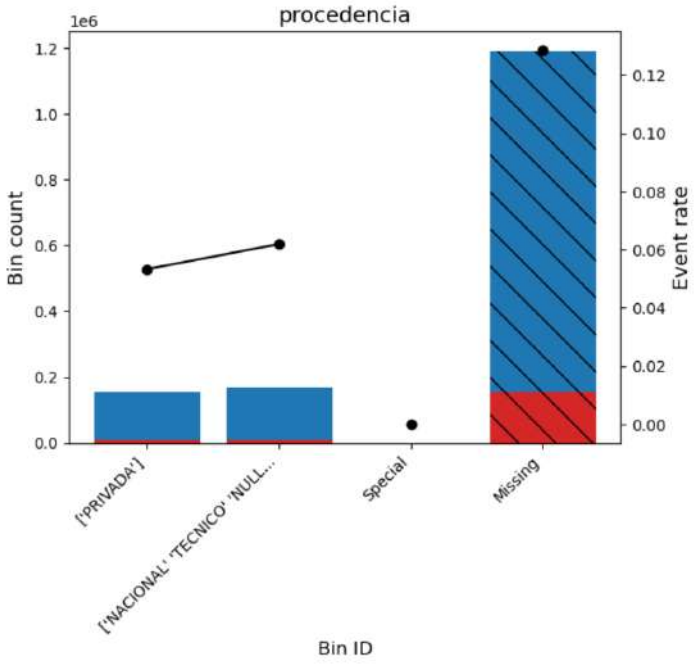
Transformación WOE para la variable Segmentación RFM

Bin	Count	Count (%)	Event rate	WoE	IV
[TOP FACTURADOR]	122661	0.08	0.09	0.19	0.0029
[DESGANADO, …]	220912	0.15	0.13	-0.14	0.003
[EN MADURACION, REENGANCHADO, …]	274648	0.18	0.14	-0.29	0.017
Missing	895878	0.59	0.10	0.11	0.007
Total	1514099	1	0.11		0.0031

Nota. Fuente: Elaboración propia

Figura 22.

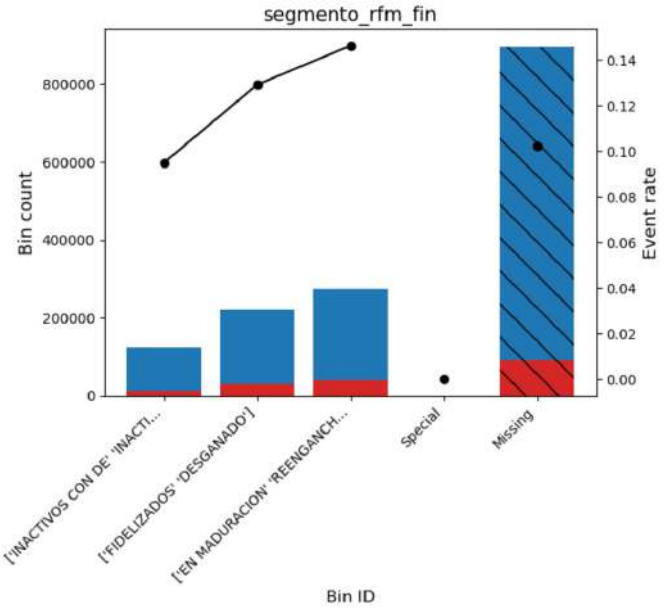
Grafico Bivariado para la variable Procedencia Profesional



Nota. Fuente: Elaboración propia

Figura 23.

Grafico Bivariado para la Variable Segmentació RFM



Nota. Fuente: Elaboración propia

El objeto OptimalBinning se guardó para su uso posterior. En la Figura 24 se observa el Optimal Binning para la variable Procedencia. Este objeto se importará más adelante para transformar variables categóricas a numéricas bajo el enfoque WOE, lo que permitirá entrenar modelos que no aceptan categorías como inputs. Una vez completado el análisis bivariado, el dataset quedó con 50 variables.

Figura 24.

Print de Objeto Optimal Binning de Procedencia

```
OptimalBinning
OptimalBinning(dtype='categorical', max_bin_size=0.95, max_n_bins=10,
               min_bin_size=0.05, min_n_bins=2, monotonic_trend='auto_asc_desc',
               name='procedencia')
```

Nota. Fuente: Elaboración propia

4.1.2.3 Filtro de Importancia Relativa mediante el Modelo de Prueba (Toy Model)

Como un filtro más, se entrenó un toy model con los siguientes hiperparámetros. Luego se mostró la importancia relativa según el peso SHAP. En la Tabla 16 se aprecia la configuración de toy model y en la Tabla 17 se observan dos variables con peso insignificante, las cuales se procede a eliminar y quedando con 48 variables.

Tabla 16.

Hiperparametros del Toy Model

Hiper Parámetro	Valor
learning_rate	0.8
objective	binary
metric	auc
max_depth	3
num_leaves	6

Nota. Fuente: Elaboración propia

Tabla 17.

TOP variables con mayor y menor importancia del Toy Model

TOP	Variable	Peso Shap	%Peso Relativo
TOP 1	pct_uso_tctot_u1m	0.301933	10.81%
TOP 2	min_pas_tot_3m	0.220334	7.93%
TOP 49	prm_trxsalud12m	0.0	0.0%
TOP 50	prm_trxeduc12m	0.0	0.0%

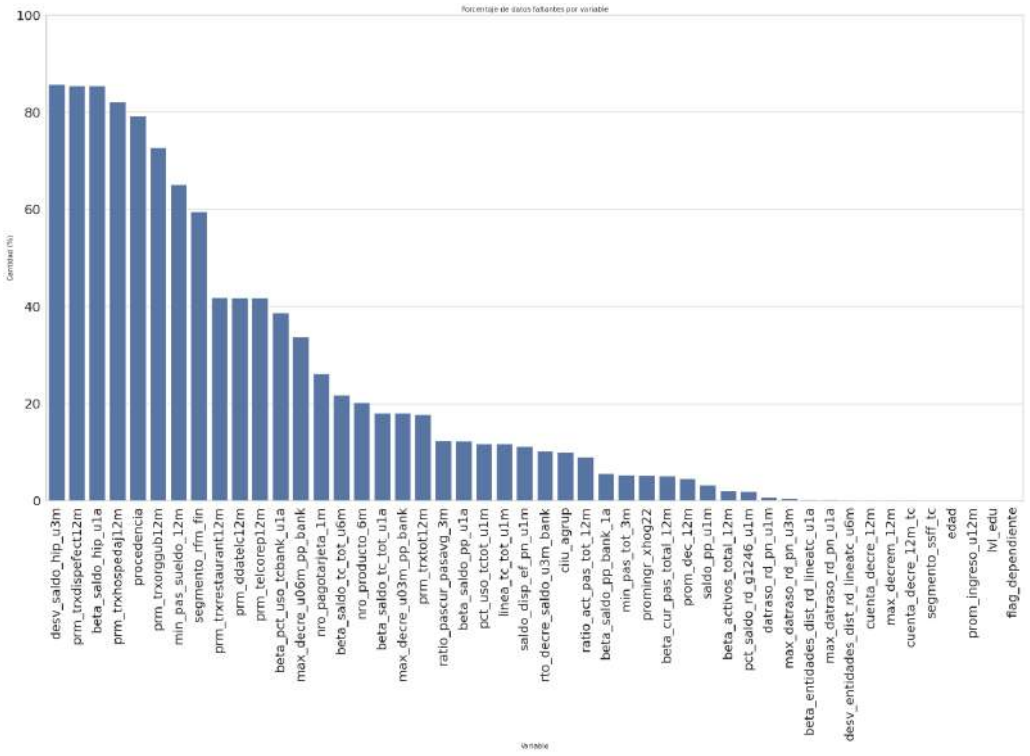
Nota. Fuente: Elaboración propia

4.1.2.4 Tratamiento de Valores Nulos

Se procedió a identificar los datos faltantes usando la librería Missingno. Como se observa en la Figura 25, algunas variables presentan un número considerable de missings, pero ninguna supera el 95 % establecido en los filtros previos.

Figura 25.

Grafico de Variables Missing calculadas con librería Missingno

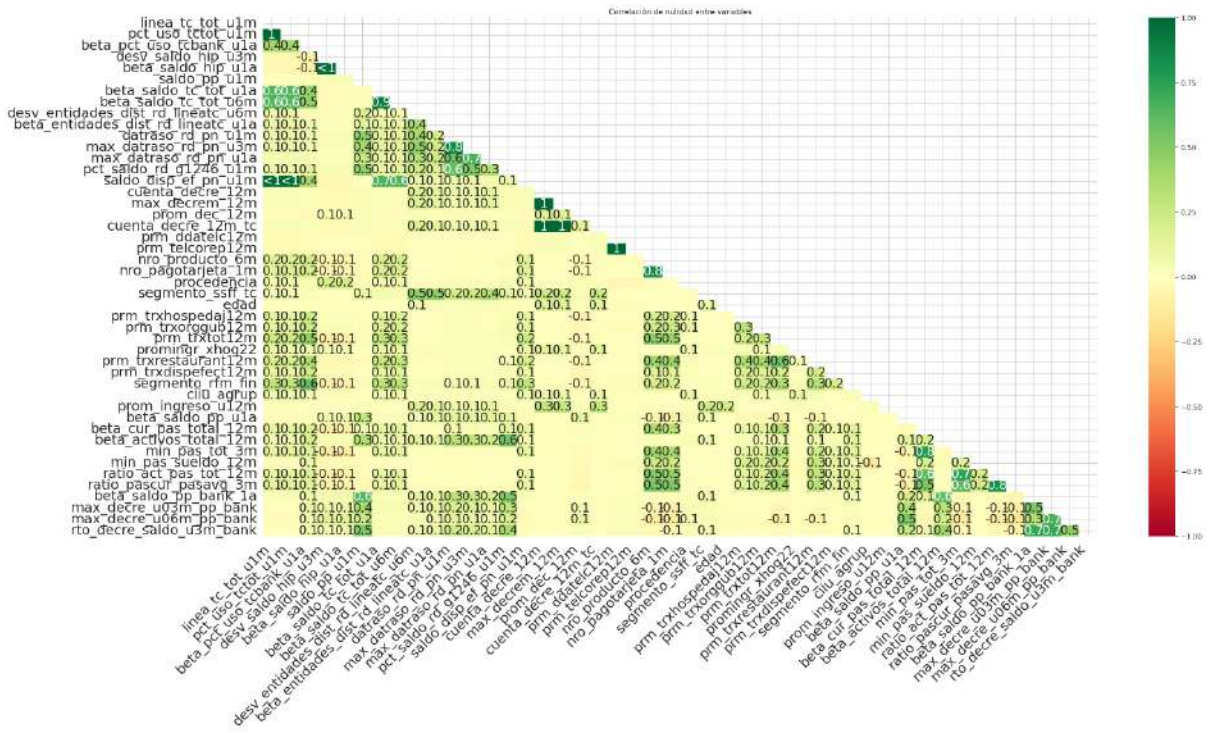


Nota. Fuente: Elaboración propia

Otro análisis importante fue verificar la relación entre los nulos. Muñoz et al. (2024) señala que existen tres tipos de missing: MCAR (aleatorio), MAR (influenciado por otra variable observable) y MNAR (influenciado por una variable no observable). El mapa de calor de la Figura 26 refleja la correlación de nulos. Algunas variables presentan una correlación cercana a 1, lo que indica que dependen de otras características. Por ejemplo, linea_tc_total_u1m tiene correlación 1 con pct_uso_tc_tot_u1m, lo cual tiene sentido ya que una persona sin tarjeta presentará nulo tanto en su línea como en su uso.

Figura 26.

Mapa de Calor Correlacional para los missing de las variables.



Nota. Fuente: Elaboración propia

El dataset presenta ausencia de nulos de tipo MAR (Missing at Random), ya que los valores faltantes no son aleatorios y están relacionados con las mismas razones por las que faltan. También presenta datos MCAR (Missing Completely at Random) en variables que no dependen de otras. Dado que el dataset contiene nulos, se imputaron todos con el valor '-999999999'. Este valor se interpretará como nulo, ya que algunos algoritmos no permiten valores faltantes durante el entrenamiento como la regresión logística. Este valor solo es referencial para poder comparar los algoritmos en los siguientes pasos.

4.1.2.5 Detección y Aislamiento de Valores Atípicos (Outliers) mediante Isolation Forest

Para la limpieza de outliers, se seleccionó el periodo 202106 – 202208 para entrenamiento y test. Sobre este conjunto, que contiene 1,514,099 registros después de los filtros (solo clientes con target Buenos y Malos), se aplicó la limpieza de datos. Se utilizó la librería sklearn.ensemble con el método Isolation Forest. Los bosques de aislamiento permiten detectar anomalías globales y locales, proporcionando datos realmente informativos y

manejando altas dimensiones e identificando anomalías complejas (Yepmo et al., 2024). El objeto Isolation Forest se instancio con los siguientes parámetros de la Tabla 18.

Tabla 18.

Parametros del Insolation Forest para al busqueda de outliers

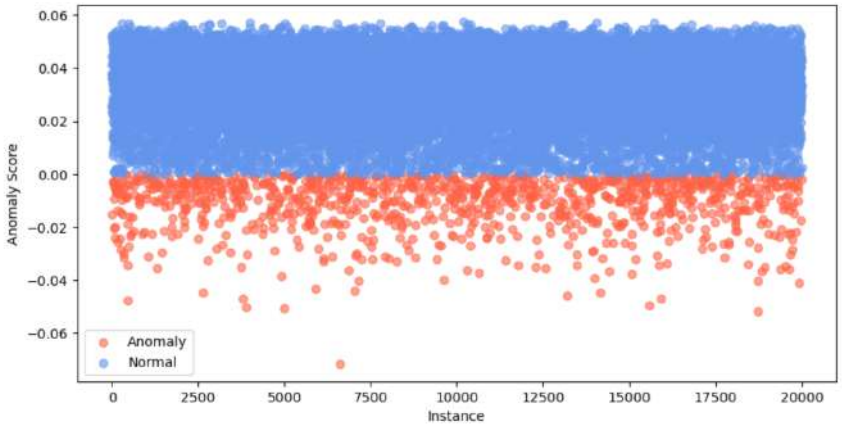
Parámetros	Explicación	Valor
n_estimators	# de árboles de aislamiento usados	100
contamination	Porcentaje que esperamos que sean outliers	0.05
Sample_size	# instancias para entrenar árbol	256
random_state	semilla	42

Nota. Fuente: Elaboración propia

En la Figura 27 se observa la visualización de anomalías por instancia para una muestra de 20,000 observaciones.

Figura 27.

Visualización de Anomalias por instancia.



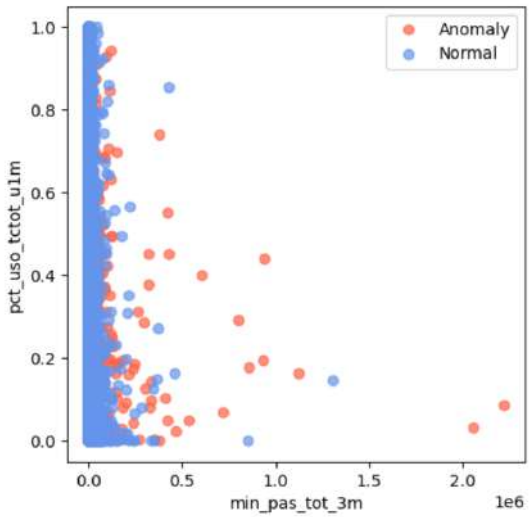
Nota. Fuente: Elaboración propia

Si comparamos las dos variables más importantes del Toy Model, pct_uso_tc_total_u1m (TOP 1) y min_pas_tot_3m (TOP 2), se puede apreciar que los valores

que escapan del gráfico de dispersión son marcados como anomalías, tal como lo muestra la Figura 28.

Figura 28.

Detección de anomalías en variables TOP

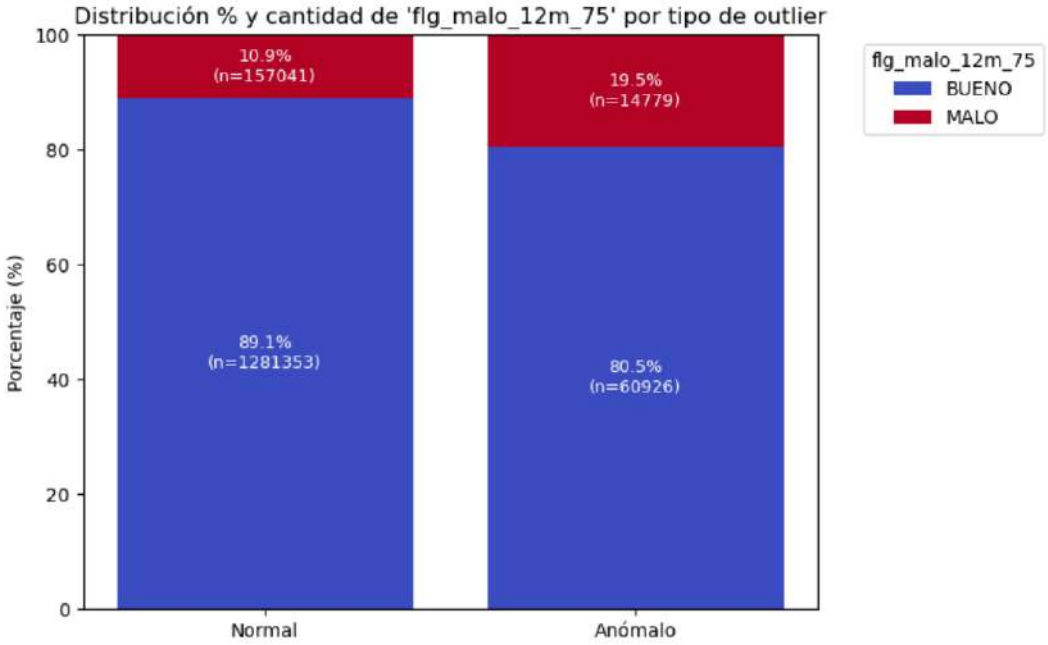


Nota. Fuente: Elaboración propia

Al observar la Figura 29, se denota la cantidad de Buenos y Malos para los registros catalogados como normales y anómalos. Las distribuciones no son totalmente diferentes en ambos grupos, de esta manera se garantiza que la eliminación de outliers no este sesgada por que el target sea Bueno o Malo.

Figura 29.

Distribución de target por registros normales y anómalos

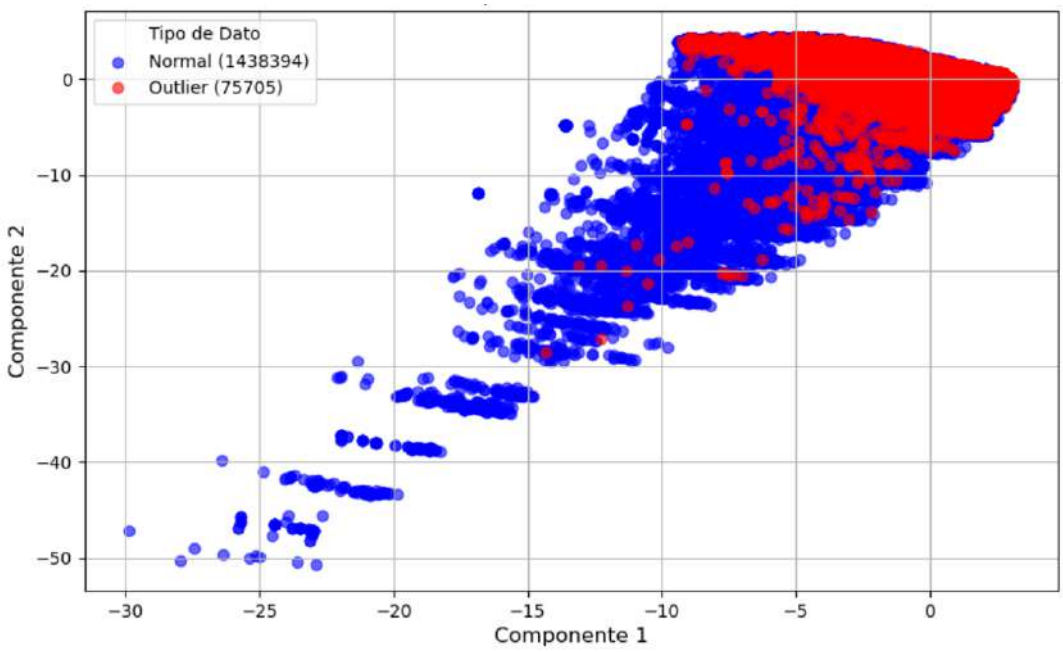


Nota. Fuente: Elaboración propia

También es fundamental poder revisar la distribución global del dataset a nivel de outliers, esto resulta difícil ya que se tiene más de 2 variables y representarlo en un espacio n-dimensional es complicado. Por esto en la Figura 30 se aprecia el resultado de aplicar PCA para visualizar en un plano 2D los registros normales vs. outliers. Visualmente, los valores outliers corresponden a datos alejados de los valores normales, lo cual nos indica que si existe una separación real entre la data normal y la atípica.

Figura 30.

Grafico Dispersión de outliers usando PCA



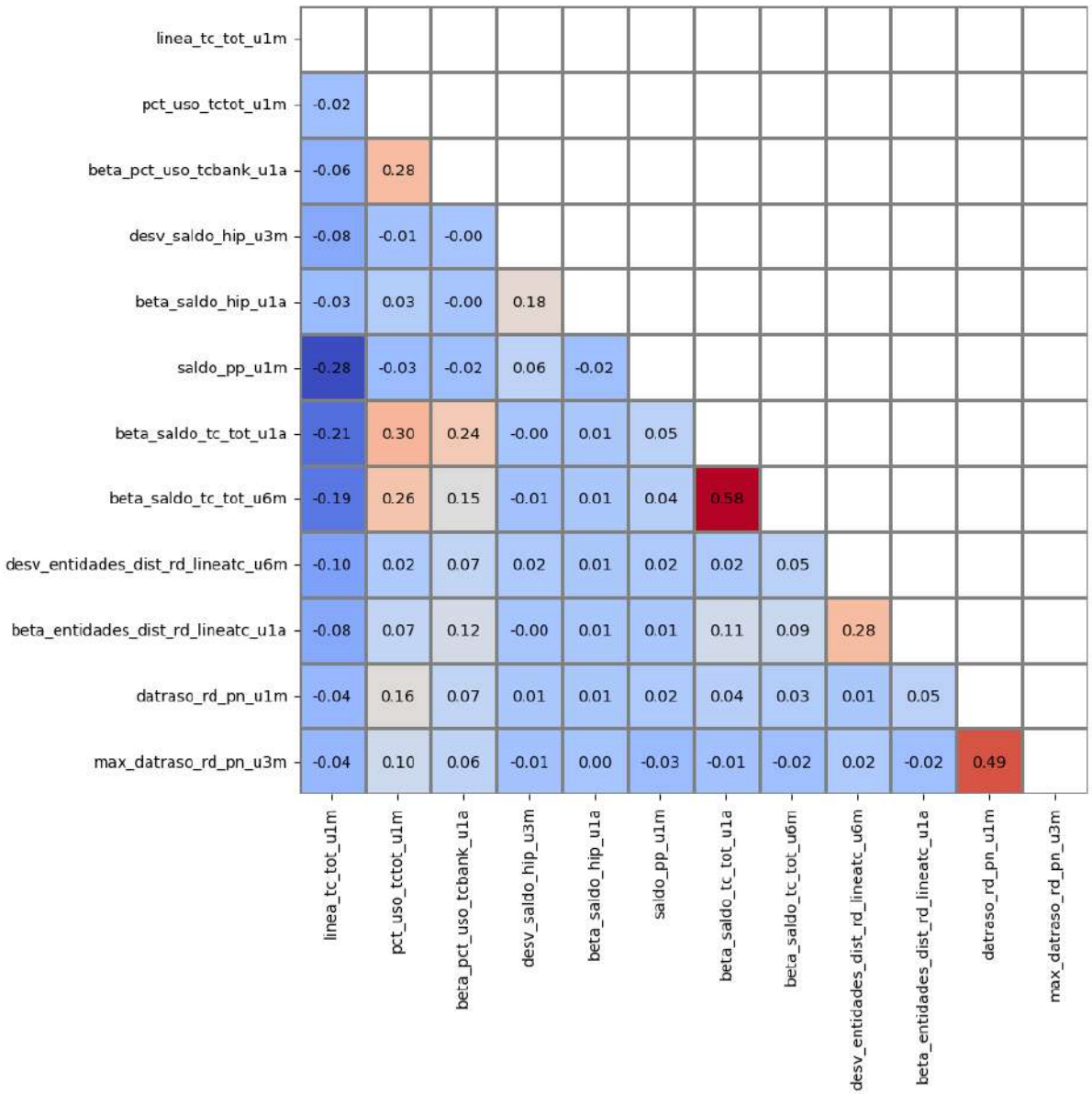
Nota. Fuente: Elaboración propia

4.1.2.6 Análisis de Correlacion del Dataset Final

En este apartado se verificó que todas las variables tuvieran una correlación menor al 60 %, tal como se estableció en secciones anteriores. Muhammad Noor et al. (2024) señala que seleccionar y filtrar características en función de su correlación mejora el rendimiento final del modelo. Si se seleccionan variables muy correlacionadas, se pierde el sentido de incluir variables con el mismo significado. En la Figura 31 se aprecia el mapa de calor de correlación para un set de las variables del dataset final.

Figura 31.

Mapa de Calor de correlacion de set de variables finalistas



Nota. Fuente: Elaboración propia

4.1.3 Evaluación, entrenamiento y selección del mejor modelo predictivo de impago para Préstamos Personales

4.1.3.1 Estrategia de Partición y Sampleo Estratificado del Dataset

Para realizar una comparativa rápida entre algoritmos, se aplicó un muestreo estratificado sobre todo el dataset de entrenamiento usando la función train_test_split. Se tomó aproximadamente el 10 % del conjunto original de entrenamiento, equivalente a 143,839 registros. Los parámetros utilizados se muestran en la Tabla 19.

Tabla 19.

Parametros del objeto train_test_split

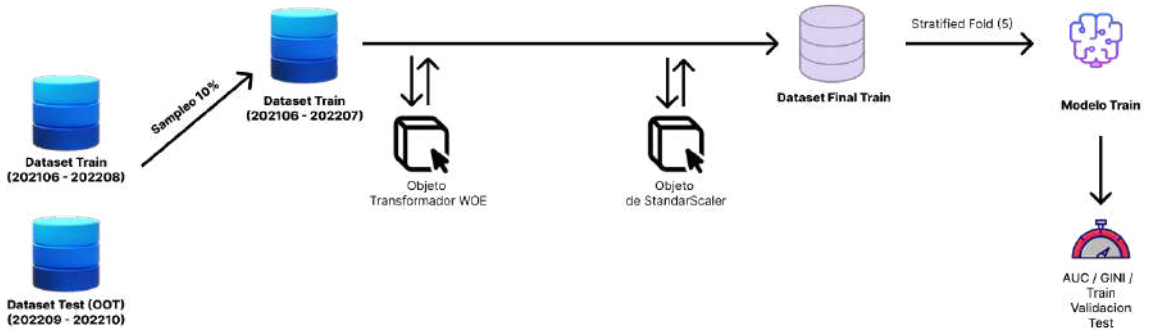
Parámetro	valor
train_size	0.1
stratify	target
random_state	42

Nota. Fuente: Elaboración propia

La Figura 32 muestra el flujo general de entrenamiento para cada modelo. El dataset de Train se somete a un muestreo estratificado del 10 %. Luego se aplican los objetos WOE y scalers previamente generados. Finalmente, el dataset de entrenamiento ingresa al flujo de entrenamiento mediante una evaluación con k-fold estratificado, calculando AUC y Gini para cada algoritmo.

Figura 32.

Flujo de entrenamiento general para benchmarking de modelos



Nota. Fuente: Elaboración propia

4.1.3.2 Entrenamiento y Validación Cruzada de Algoritmos Candidatos

En esta sección se detalla el proceso de modelamiento predictivo. Se entrenaron y evaluaron múltiples algoritmos para establecer una línea base de rendimiento (baseline). El objetivo central es seleccionar el modelo óptimo que responda al problema de estimación planteado.

4.1.3.2.1 Regresión Logística

El primer algoritmo entrenado fue la regresión logística, este algoritmo ofrece mayor interpretabilidad en comparación con otras alternativas. La regresión logística permite conocer la fórmula y los pesos de cada variable, esto se puede ver en la Ecuación 2.

Ecuación 2:

Regresión Logistica

$$\log\left(\frac{p}{1-p}\right) = b_0 + \sum_{1}^{k} b_k X_k \quad (2)$$

Donde:

- p = Probabilidad de ocurrencia del evento
- b_0 = Intercepto
- b_k = Coeficiente del predictor
- X_k = Variables Predictoras

El entrenamiento y la validación cruzada se realizaron con StratifiedKFold usando los parámetros de la Tabla 20, en donde también se puede ver los parámetros del modelo.

Tabla 20.

Parametros Del StratifiedKFold para la Regresión Logística

Parámetro	Explicación	Valor
n_splits	Numero de pliegues	5
max_iter	Valores grandes cerca de 1000 cuando se usa algunos solver como sag	1000
solver	Stochastic Average Gradient Eficiente para datasets grandes, alta dimensionalidad	sag

Nota. Fuente: Elaboración propia

Los resultados del entrenamiento se muestran en la Tabla 21. El modelo obtuvo un Gini máximo de 0.38 en el conjunto de test.

Tabla 21.

Resultados del Kfold para la Regresión Logística

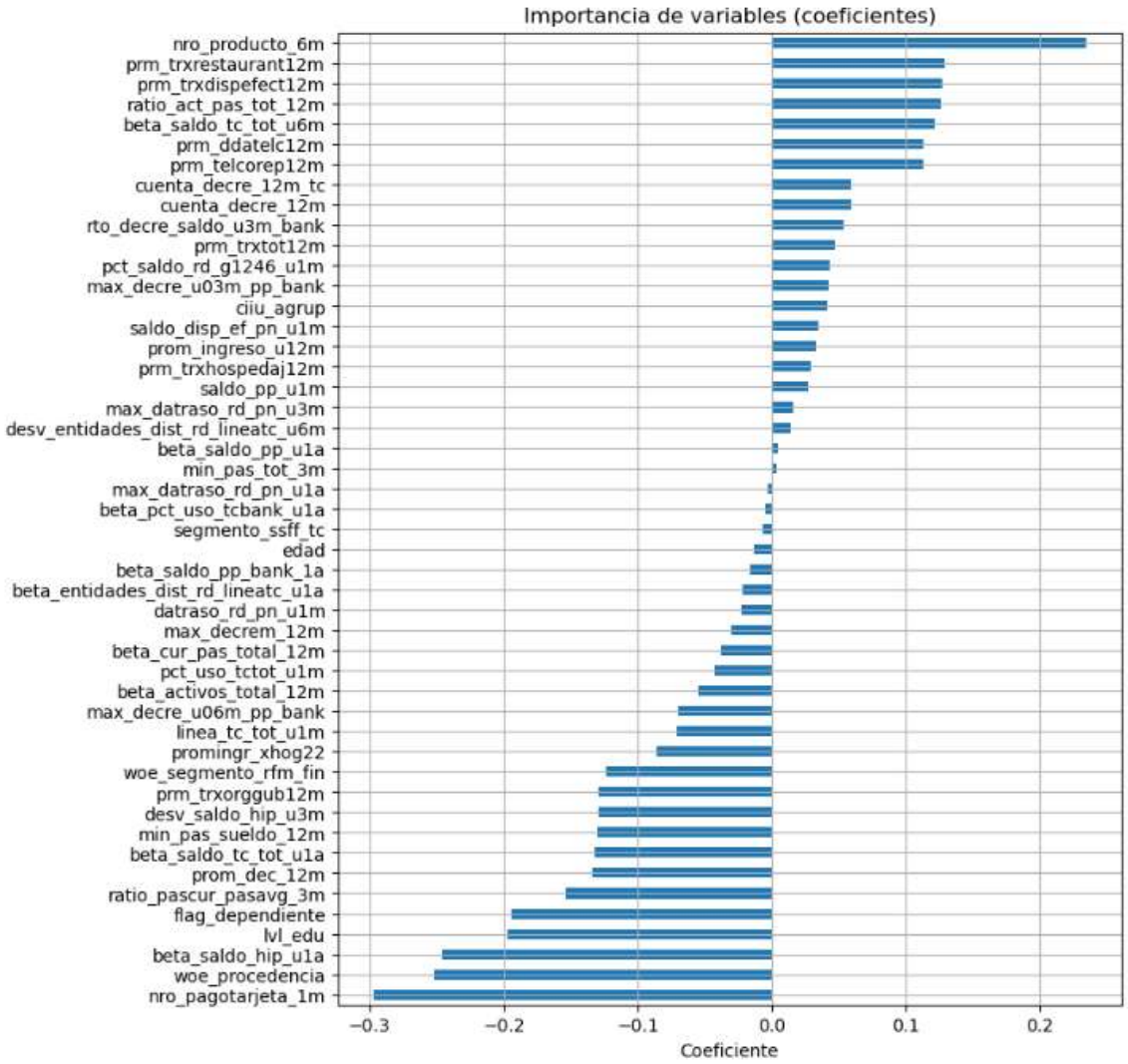
# Fold	Gini Validacion	Gini Test
Fold 1	0.37	0.38
Fold 2	0.39	0.37
Fold 3	0.39	0.38
Fold 4	0.37	0.37
Fold 5	0.37	0.37

Nota. Fuente: Elaboración propia

Se seleccionó el mejor modelo y se extrajeron los coeficientes. La variable top 1 fue nro_pagotarjeta_1m, con un coeficiente de -0.29. En términos de odds ratio, un incremento de una unidad en esta variable reduce las odds del evento en aproximadamente un 25 % (odds ratio = $e^{-0.29} \approx 0.75$). La variable nr_producto_6m obtuvo un coeficiente de 0.23: por cada producto adicional que el cliente haya tenido en los últimos 6 meses, las odds de que ocurra el evento aumentan en un 26.4 %. La Figura 33 muestra los coeficientes del modelo.

Figura 33.

Coeficientes del Algoritmo Regresión Logística

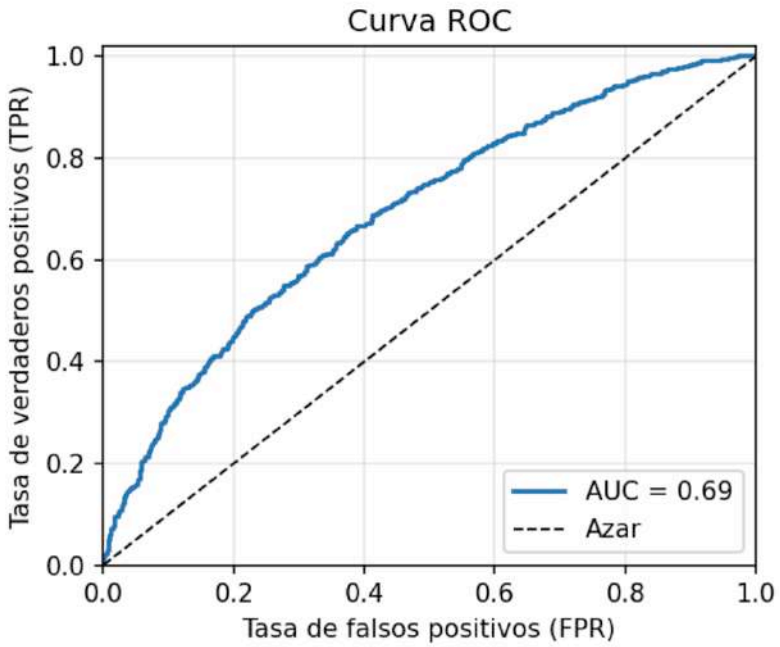


Nota. Fuente: Elaboración propia

En la Figura 34 se aprecia la curva ROC, con un AUC de 0.69, un valor muy bajo para la industria. Esto indica que el modelo tiene un poder predictivo muy bajo. Por lo tanto, no es un buen candidato para ser el modelo final.

Figura 34.

Curva ROC para el Algoritmo Regresión Logística



Nota. Fuente: Elaboración propia

4.1.3.2.2 Árbol de Decisión

De manera similar, se entrenó el algoritmo de árboles de decisión usando DecisionTreeClassifier con k-fold estratificado. Los árboles de decisión son algoritmos que presentan reglas de decisión para problemas de clasificación. Los hiperparámetros usados se muestran en la Tabla 22.

Tabla 22.

Parametros del Kfold para Algoritmo Árbol de Decisión

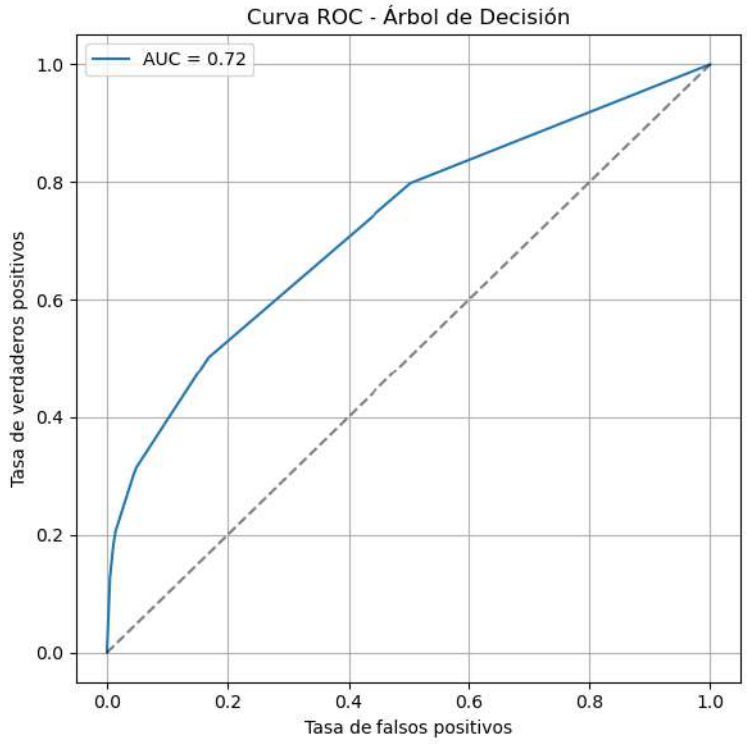
Parámetro	Explicación	Valor
#K Folds	Criterio de división	5
max_depth	Profundidad del árbol	3
min_samples_leaf	Mínimo de ejemplos para un nodo hoja	5

Nota. Fuente: Elaboración propia

Los resultados obtenidos fueron: AUC 0.71 y Gini 0.43 en validación y test. En la Figura 35 se muestra la curva ROC para la muestra de test. Este AUC se cataloga como bueno, indicando que el modelo es útil.

Figura 35.

Curva ROC para el Algoritmo Árbol de Decisión

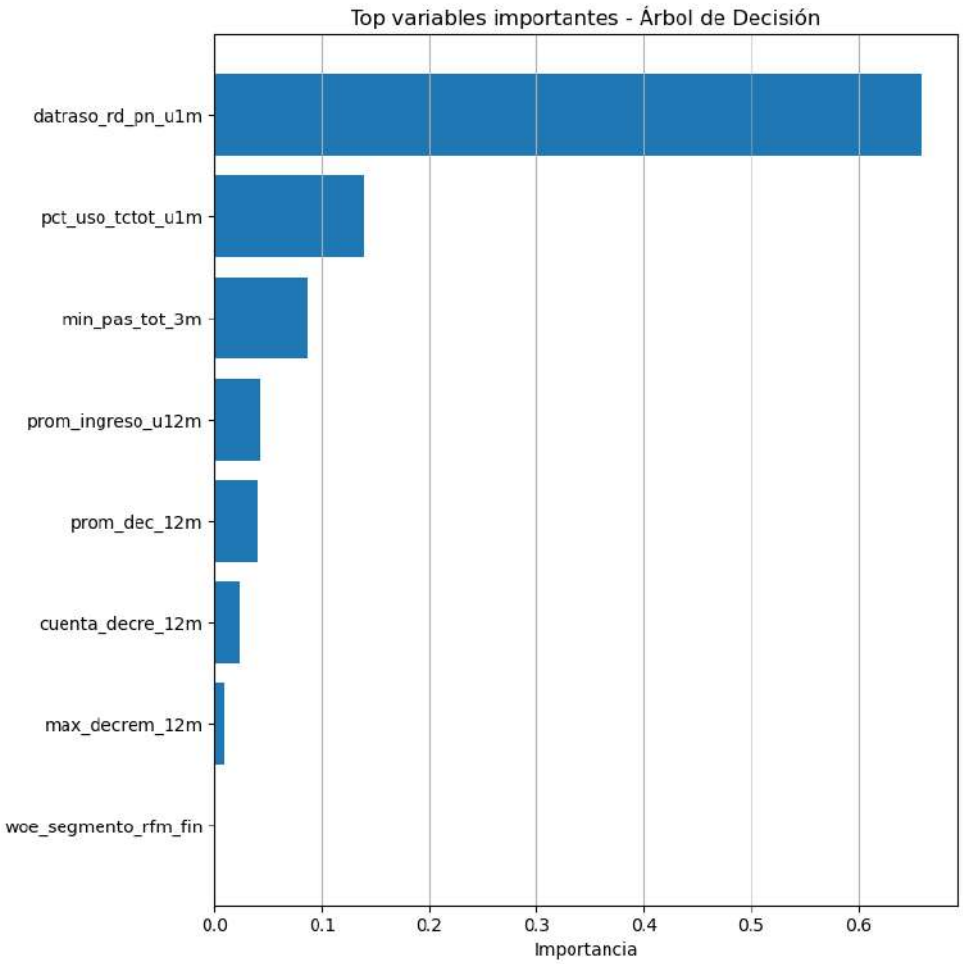


Nota. Fuente: Elaboración propia

La Figura 36 muestra las variables más importantes. A partir de la variable 8, el aporte en términos de importancia es nulo, por lo que el modelo solo toma como relevantes las 7 primeras variables.

Figura 36.

Top Variables mas importanes del Árbol de decisión



Nota. Fuente: Elaboración propia

4.1.3.2.3 Random Forest

Los modelos de Random Forest se utilizan ampliamente en clasificación supervisada. Son útiles en grandes conjuntos de datos y permiten capturar relaciones no lineales. Los hiperparámetros usados se muestran en la Tabla 23.

Tabla 23.

Parametros para Algoritmo Random Forest

Parámetro	Explicación	Valor
n_estimators	Número de árboles	100
max_depth	Profundidad máxima de cada árbol	5
min_samples_leaf	Mínimo de muestras en una hoja	10

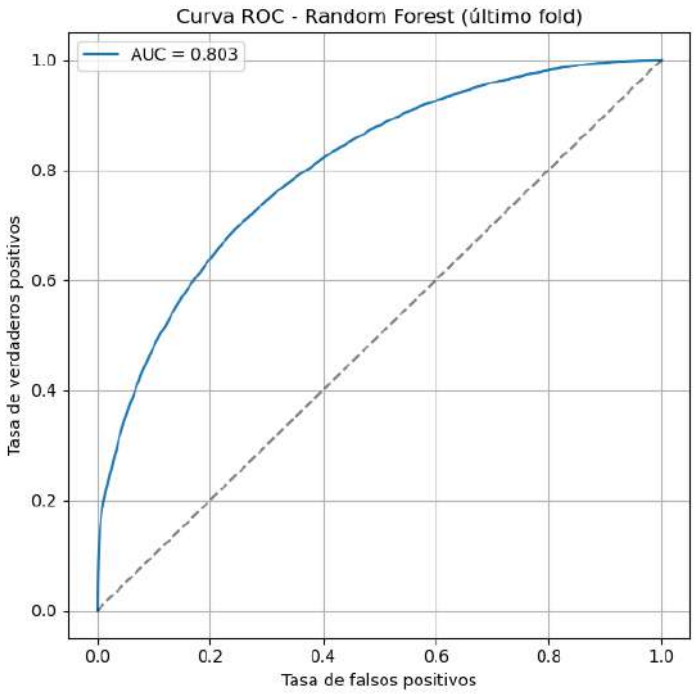
random_state	Semilla de aleatoriedad	42
n_jobs	CPUs en paralelo	-1

Nota. Fuente: Elaboración propia

En la Figura 37 se aprecia que Random Forest logra un AUC de 0.8 y un Gini de 0.61 en validación. Este Gini supera a los dos modelos anteriores debido a la naturaleza del algoritmo, la eficacia del Random Forest permite predecir mejor en comparación con otros modelos.

Figura 37.

Curva ROC para el Algoritmo Random Forest

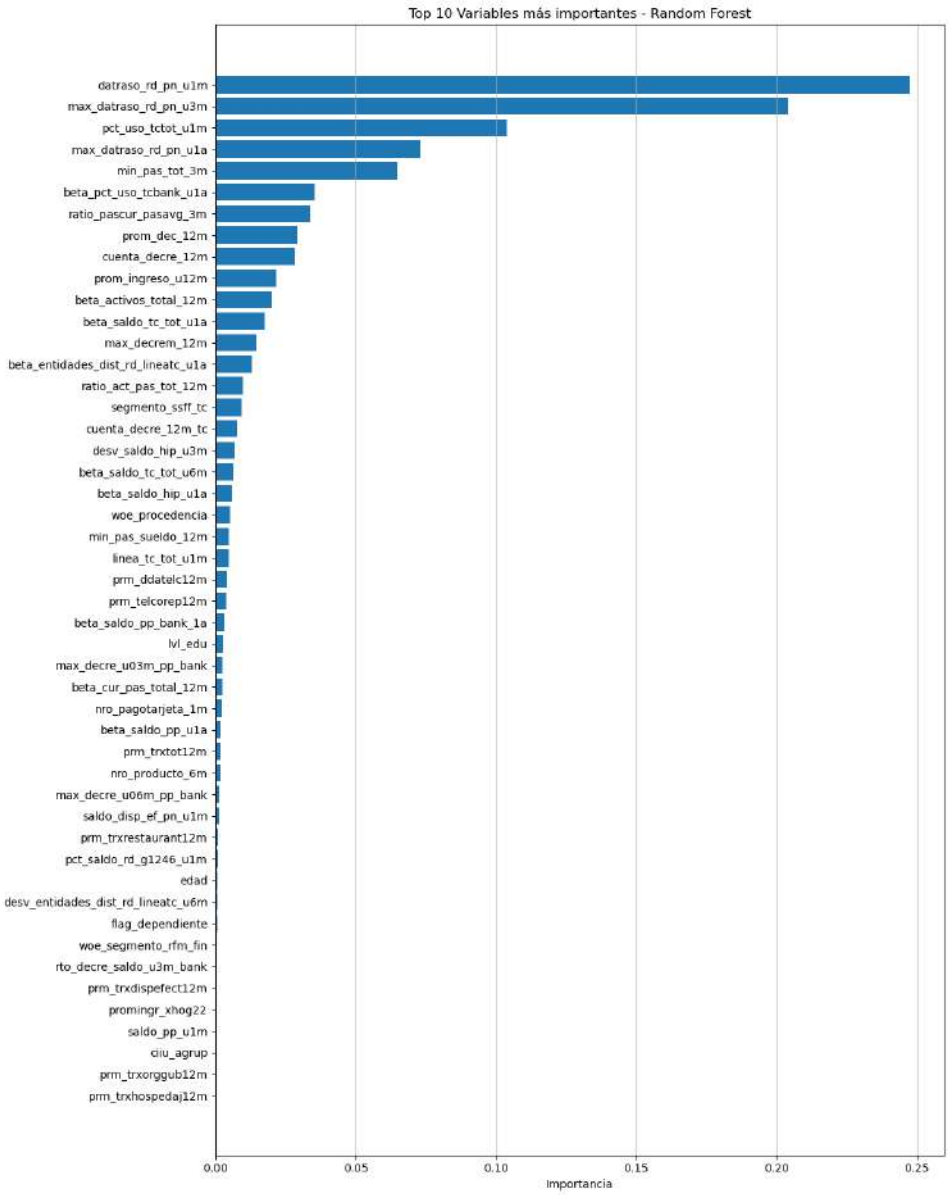


Nota. Fuente: Elaboración propia

Este algoritmo distribuye mejor la importancia de las variables al momento de hacer la inferencia. En la Figura 38 se observa que Random Forest asigna pesos de manera más equilibrada.

Figura 38.

Top Variables para el Algoritmo Random Forest



Nota. Fuente: Elaboración propia

4.1.3.2.4 AdaBoost Classifier

AdaBoost es un modelo de boosting basado en la combinación de árboles débiles, este algoritmo entrena múltiples clasificadores débiles y se combinan sus resultados para obtener una clasificación más precisa y suele ser más predictivo. AdaBoost logra superar a Random

Forest y SVM en varios casos, debido a la creación de weak learners que mejoran en cada ronda a partir de errores anteriores. Los hiperparámetros usados se muestran en la Tabla 24.

Tabla 24.

Parametros para Algoritmo AdaBoost Classifier

Parámetro	Explicación	Valor
n_estimators	Número de árboles débiles	100
learning_rate	Contribución de cada modelo	1.0
base_estimator	Árbol débil con max_depth = 1	1

Nota. Fuente: Elaboración propia

Los resultados se muestran en la Tabla 25. En Train se obtuvo un Gini de 63.2 %, superando a los modelos anteriores. En validación y test se obtuvo 0.622 y 0.623 respectivamente. Este indicador se considera muy bueno para predicción y reflejo de ello la Figura 39 muestra la curva ROC.

Tabla 25.

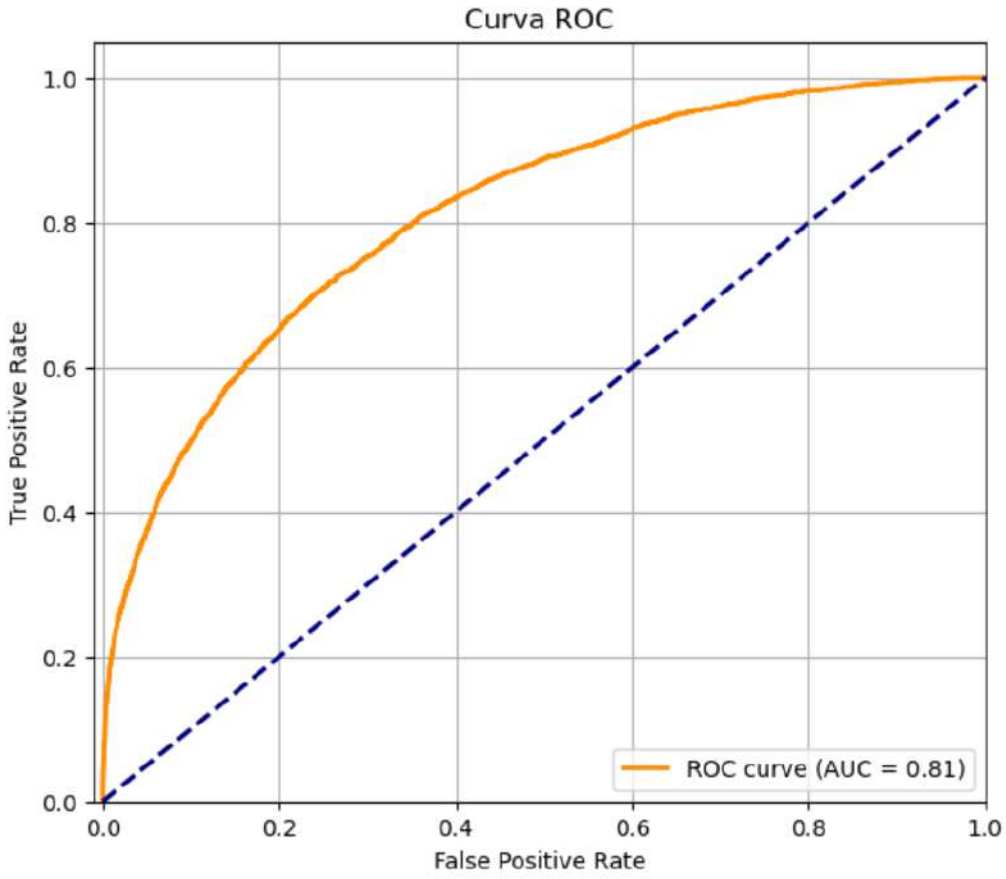
Resultados KFold para Algoritmo AdaBoost Classifier

Parámetro	AUC	Gini
Train	0.816	0.632
Validación	0.811	0.622
Test	0.812	0.623

Nota. Fuente: Elaboración propia

Figura 39.

Curva ROC para el Algoritmo AdaBoost Clasifier

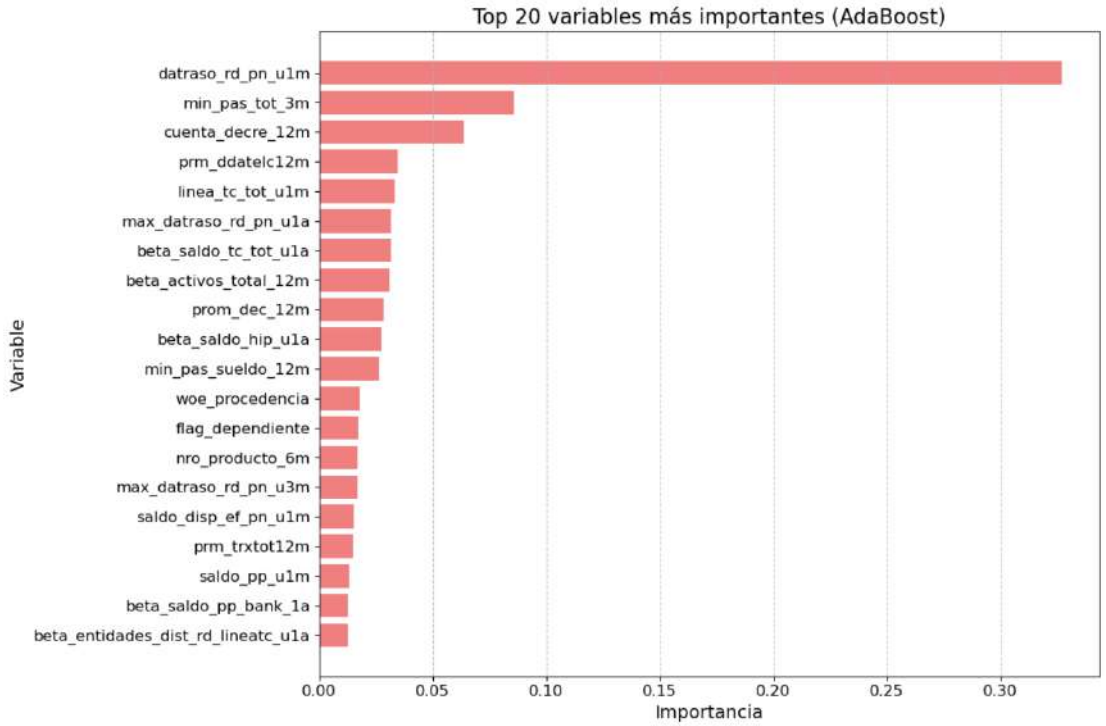


Nota. Fuente: Elaboración propia

En la Figura 40 se observa el top 20 de variables más importantes. Se aprecia que un mayor número de variables toma importancia en el modelo debido a su naturaleza de Boosting.

Figura 40.

Top Variables para el Algoritmo AdaBoost Clasifier



Nota. Fuente: Elaboración propia

4.1.3.2.5 CatBoost Classifier

En el mundo bancario, CatBoost es muy utilizado, este algoritmo permite técnicas de regularización para evitar sobreajuste y mejorar la generalización. CatBoost forma parte de la familia de boosting y destaca por la posibilidad de introducir variables categóricas sin tratamiento previo. CatBoost suele ser más predictivo que AdaBoost, ya que usa árboles más complejos y profundos, sumado a técnicas como target encoding. Se encontró una mejora de aproximadamente 5 puntos de precisión frente a AdaBoost en predicción de contaminación del aire (Naidu & Amanullah, 2025). Los hiperparámetros usados se muestran en la Tabla 26.

Tabla 26.

Parametros para Algoritmo CatBoost Classifier

Parámetro	Explicación	Valor
iterations	Número de árboles	100
learning_rate	Peso de cada árbol	0.1
depth	Profundidad de cada árbol	5

eval_metric	Métrica de evaluación	AUC
l2_leaf_reg	Regularización L2	3
random_strength	Selección de splits	1
random_seed	Semilla de aleatoriedad	42

Nota. Fuente: Elaboración propia

Los resultados se muestran en la Tabla 27. CatBoost obtuvo un Gini de 64.6 % en test, posicionándose como el mejor hasta ese momento. La curva ROC de la Figura 41 confirma el buen poder predictivo.

Tabla 27.

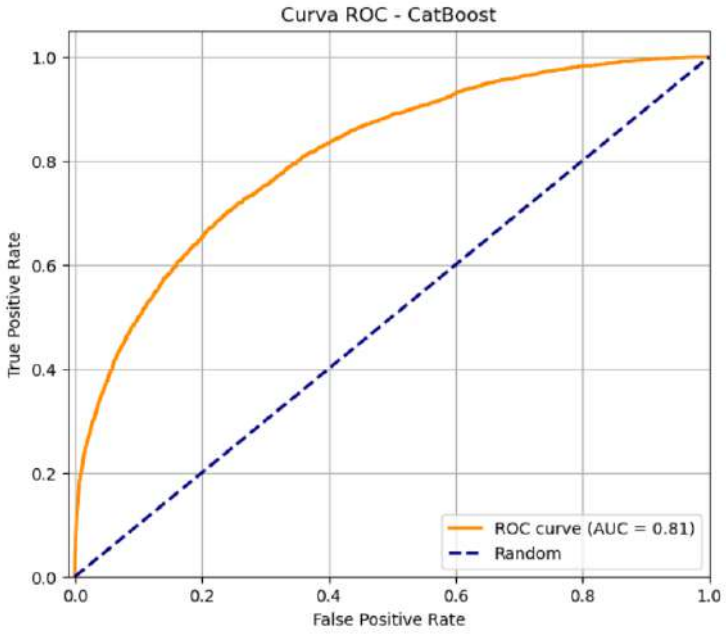
Resultados KFold para Algoritmo CatBoost Classifier

Parámetro	AUC	Gini
Train	0.833	0.666
Validación	0.824	0.647
Test	0.823	0.646

Nota. Fuente: Elaboración propia

Figura 41.

Curva ROC para el Algoritmo CatBoost Clasifier

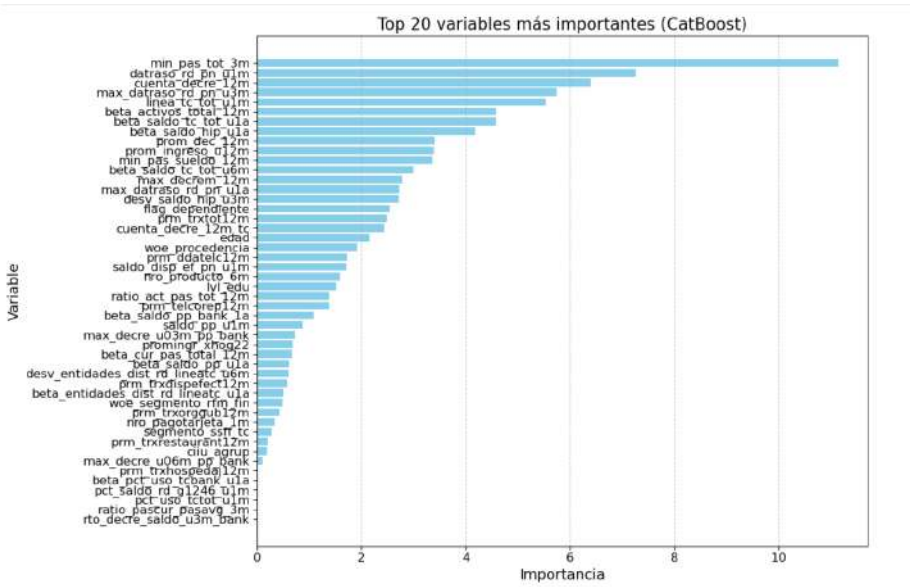


Nota. Fuente: Elaboración propia

En la Figura 42 se observa una mejor distribución de pesos entre las variables.

Figura 42.

Top Variables para el Algoritmo CatBoost Clasifier



Nota. Fuente: Elaboración propia

4.1.3.2.6 LightGBM Classifier

LightGBM es uno de los algoritmos más famosos de la familia boosting. Paz (2024) señala que se caracteriza por su eficiencia computacional, manejo de datasets pesados y alto rendimiento predictivo. Este algoritmo es muy usado en inteligencia artificial aplicada al sector bancario, LGBM muestra un gran balance en métricas generales, posicionándolo como una buena alternativa para la gestión financiera moderna. Los hiperparámetros usados se muestran en la Tabla 28.

Tabla 28.

Parametros para Algoritmo LightGBM Classifier

Parámetro	Explicación	Valor
iterations	Número de árboles	100

learning_rate	Peso de cada árbol	0.08
max_depth	Profundidad de cada árbol	4
eval_metric	Métrica de evaluación	AUC
objective	Tipo de tarea	binary
boosting_type	Tipo de gradiente	gbdt
colsample_bytree	Fracción de features usadas	1.0

Nota. Fuente: Elaboración propia

Los resultados se muestran en la Tabla 29 y la Figura 43 refleja la Curva ROC. LightGBM obtuvo un Gini de 68.9 % en test, el más alto hasta el momento. Sin embargo, se observa una diferencia importante entre train, validación y test.

Tabla 29.

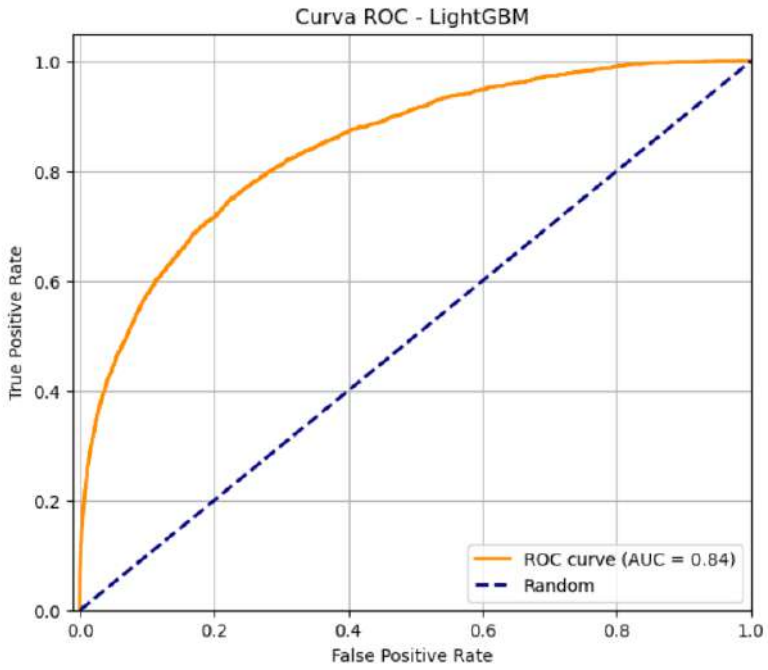
Resultados para Algoritmo LightGBM Classifier

Parámetro	AUC	Gini
Train	0.858	0.716
Validación	0.835	0.669
Test	0.845	0.689

Nota. Fuente: Elaboración propia

Figura 43.

Curva ROC para el Algoritmo LightGBM Classifier

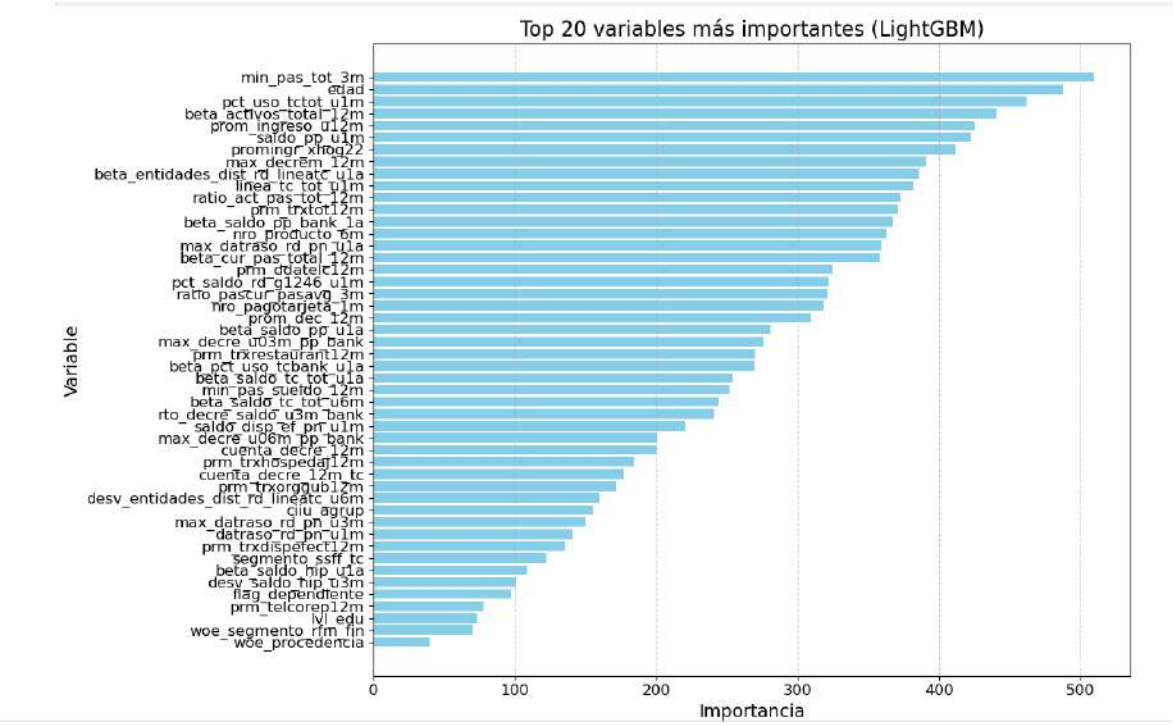


Nota. Fuente: Elaboración propia

Este algoritmo, al detectar tendencias no lineales, permite una mejor distribución de las variables importantes y tiene la capacidad de medir relaciones no lineales. La Figura 44 muestra el peso de las variables más importantes.

Figura 44.

Peso de Variables para el Algoritmo LightGBM Clasifier



Nota. Fuente: Elaboración propia

4.1.3.2.7 XGBOOST Classifier

XGBoost es un modelo muy usado para clasificación binaria. Es versátil, diseñado para ser veloz y confiable, con un amplio abanico de aplicaciones. Ofrece un manejo extenso de hiperparámetros y un tuneo muy específico. Los hiperparámetros usados se muestran en la Tabla 30.

Tabla 30.

Parametros para Algoritmo XGBoost Classifier

Parámetro	Explicación	Valor
iterations	Número de árboles	100
learning_rate	Peso de cada árbol	0.08
max_depth	Profundidad de cada árbol	5
eval_metric	Métrica de evaluación	AUC
objective	Tipo de tarea	binary

Nota. Fuente: Elaboración propia

XGBoost obtuvo un Gini de 76.3 % en train, pero en test fue de 66.2 %, mostrando una diferencia de casi 10 puntos porcentuales, lo que indica cierto sobreajuste tal como se ve en la Tabla 31 y en la Figura 45. La Figura 46 muestra el Top de variables más importantes.

Tabla 31.

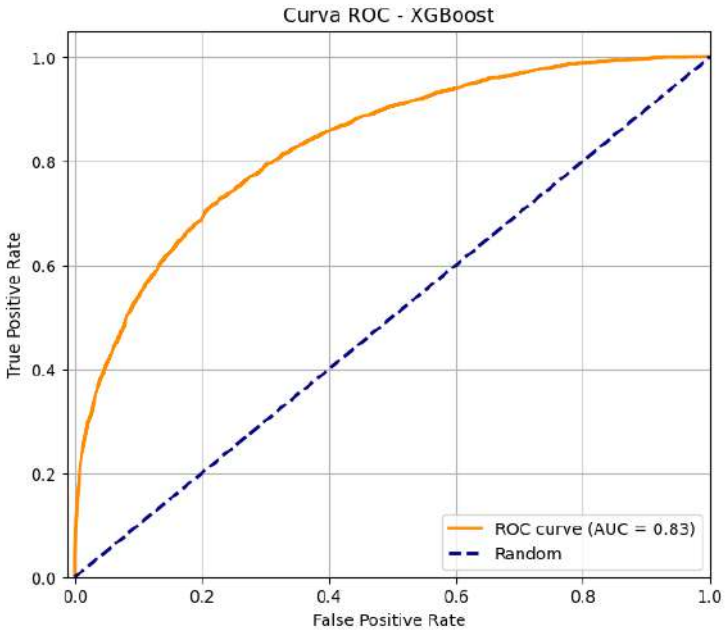
Resultados para Algoritmo XGBoost Classifier

Parámetro	AUC	Gini
Train	0.881	0.763
Validación	0.830	0.661
Test	0.831	0.662

Nota. Fuente: Elaboración propia

Figura 45.

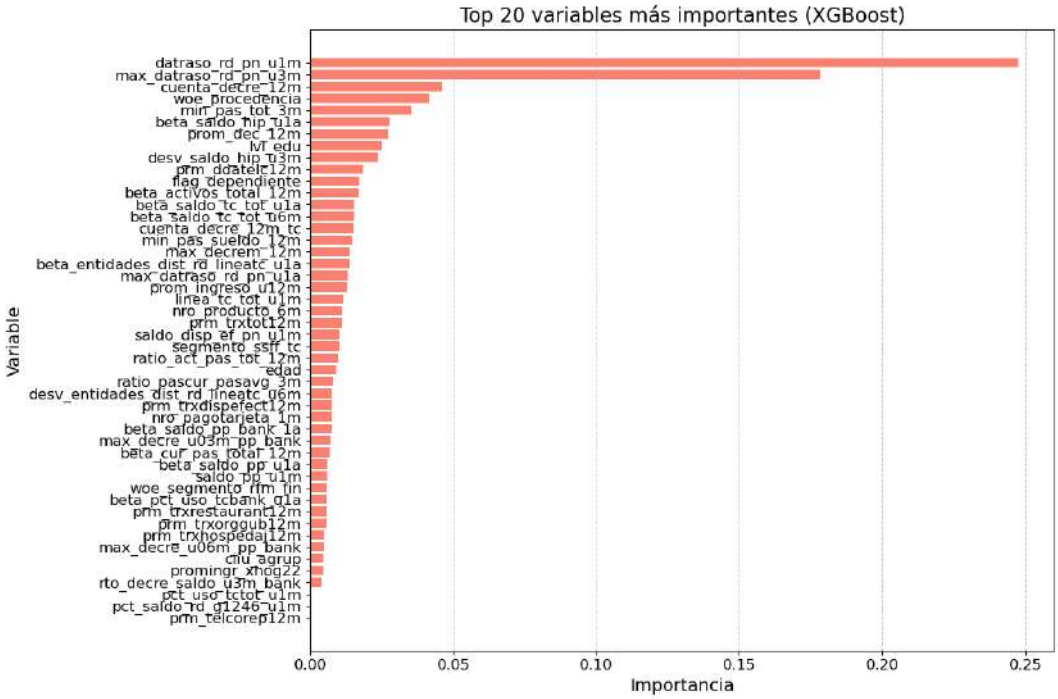
Curva ROC para el Algoritmo XGBoost Classifier



Nota. Fuente: Elaboración propia

Figura 46.

TOP Variables mas importantes para el Algoritmo XGBoost Classifier



Nota. Fuente: Elaboración propia

4.1.3.3 Comparativa de Modelos Candidatos (Benchmarking)

Se entrenaron 7 algoritmos sin optimización de hiperparámetros. En esta sección se resumen las métricas principales para seleccionar el mejor candidato, la Tabla 32 muestra un resumen de los resultados obtenidos.

Tabla 32.

Resumen de Resultados obtenidos por cada modelo.

Modelo	GINI máximo (train/valid)	GINI test	AUC test	Diferencia Gini
XGBoost	76.3 %	66.2 %	83.1 %	-10.1 %
LightGBM	71.6 %	68.9 %	84.5 %	-2.7 %
CatBoost	66.6 %	64.6 %	82.3 %	-2.0 %
AdaBoost	63.2 %	62.3 %	81.2 %	-0.9 %
Random Forest	61.0 %	61.0 %	80.0 %	0.0 %
Árbol de Decisión	43.0 %	43.0 %	71.0 %	0.0 %

Regresión Logística	39.0 %	38.0 %	69.0 %	-1.0 %

Nota. Fuente: Elaboración propia

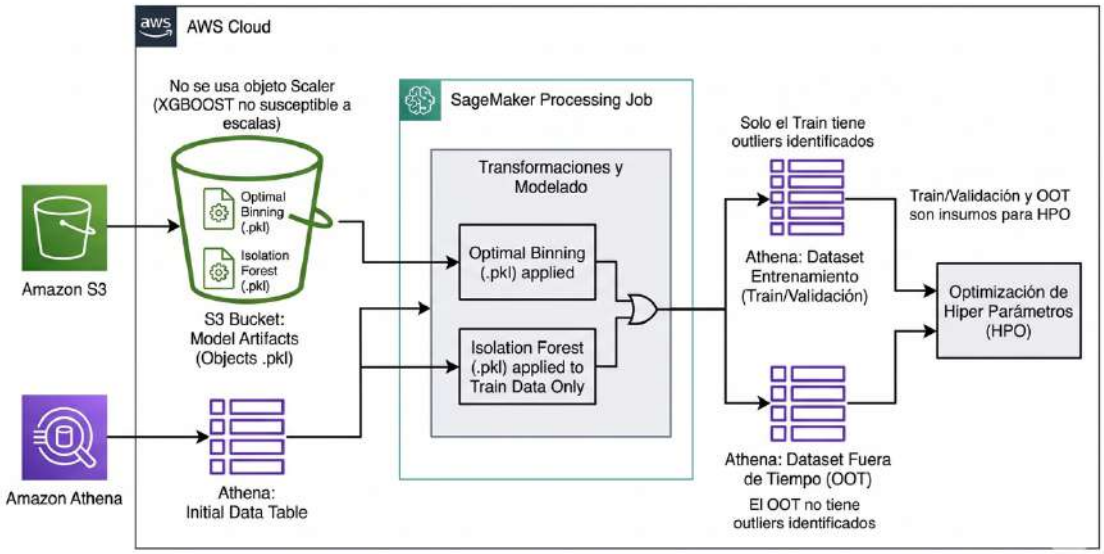
XGBoost obtuvo el mayor rendimiento en train y el segundo en test, pero con una diferencia significativa que indica sobreajuste. Yin et al. (2024) mencionan que XGBoost permite un manejo eficiente de datos desequilibrados y ofrece formas de regularización para evitar el sobreajuste. El uso de estos hiperparámetros extras será clave para reducir el sobreajuste. El modelo seleccionado es XGBoost, por su mayor rendimiento en train lo cual indica que tiene mayor potencial de mejora. El problema del sobreajuste se tratará mediante optimización de hiperparámetros y manejo del desbalance de clases.

4.1.3.4 Optimización Bayesiana de Hiperparámetros mediante Arquitectura Cloud en Amazon SageMaker

Se utilizó la librería Optuna para optimizar hiperparámetros. Esta librería permite definir una función objetivo a maximizar o minimizar. El proceso sigue el concepto de optimización bayesiana bajo el algoritmo TPE, que busca combinaciones óptimas en función de la métrica objetivo. Das et al. (2020) indican que Amazon Web Services permite realizar optimización de hiperparámetros en Amazon SageMaker mediante Processing Jobs. Se utilizó esta herramienta para aprovechar el poder computacional. En la Figura 47 se aprecia el flujo del Processing Job. Se consulta un bucket S3 donde se encuentran los objetos PKL del Isolation Forest y el Optimal Binning. No se usó el objeto Scaler porque XGBoost no es susceptible a valores en diferentes escalas. El servicio consulta una tabla de Athena con el dataset inicial y aplica las transformaciones. Finalmente, el dataset de entrenamiento y el de fuera de tiempo (OOT) son insumos para la optimización.

Figura 47.

Diagrama de Arquitectura para el flujo de ejecucion y optimizacion del Modelo



Nota. Fuente: Elaboración propia

La data usada para entrenamiento ya no es una muestra del 10 %. En la Tabla 33 se aprecia que el periodo de entrenamiento abarca desde 202106 hasta 202208, usando k-fold estratificado = 5. El periodo OOT comprende 202209 al 202210, para obtener la métrica oficial. No se excluyeron los outliers de la muestra OOT para probar la robustez.

Tabla 33.

Detalle de Periodos Usados en Entrenamiento y OOT

Data	Periodos	Outliers	# Registros	Total
202106-202208	Train	Sí	1,150,715	1,438,394
	Validación	Sí	287,679	
202209-202210	Test (OOT)	No	221,241	221,241

Nota. Fuente: Elaboración propia

La Tabla 34 muestra los atributos enviados a la función objetivo. Se buscó optimizar el AUC, con una learning rate entre 0.01 y 0.1, max_depth entre 2 y 6, y max_leaves entre 2 y 16. Se trató el sobreajuste con gamma, lambda y alpha. También se

envió scale_pos_weight para controlar el desbalance, calculado como la Ecuación 3 dando un resultado de aproximadamente 8.15.

Ecuación 3:

Scale Pos Weight

$$scale_pos_{weight} = \frac{negativos}{positivos} \quad (3)$$

Donde:

- $negativos$ = cantidad de clientes buenos
- $positivos$ = cantidad de clientes malos

Tabla 34.

Espacio de Búsqueda y Valor Encontrado para la Optimización de Hiperparametros

Parámetro	Espacio de Búsqueda	Valor Encontrado	Descripción práctica para Negocio
objective	binary:logistic	binary:logistic	Clasificar en dos grupos devolviendo probabilidad
eval_metric	auc	auc	Mide qué tan bien separa casos positivos y negativos
verbosity	2	2	Nivel de detalle técnico
tree_method	hist	hist	Método de procesamiento rápido
scale_pos_weight	8.15	8.15	Ajuste para equilibrar datos desbalanceados
eta (learning_rate)	0.01 – 0.1	0.09	Ritmo de aprendizaje

subsample	0.5 - 1	0.61	Porcentaje de datos usado para cada árbol
colsample_bytree	0.5 - 1	0.79	Fracción de variables elegidas al azar
colsample_bylevel	0.5 - 1	0.6	Fracción de variables elegidas en cada nivel
max_depth	2 - 4	4	Límite de complejidad del árbol
max_leaves	2 - 10	10	Número máximo de nodos finales
min_child_weight	1 - 50	45	Cantidad mínima de evidencia para crear una regla
max_delta_step	0 - 10	4	Estabiliza el modelo en datos desbalanceados
gamma	0 - 5	0.85	Requisito de mejora mínima para crecer
lambda (reg_lambda)	0 - 10	0.004	Penalización para evitar dar importancia a variables poco útiles
alpha (reg_alpha)	0 - 10	0.55	Penalización que elimina variables ruidosas

Nota. Fuente: Elaboración propia

4.1.3.5 Evaluación de Performance del Modelo Final (Métricas ROC, Gini, KS y Matriz de Confusión)

Las métricas del modelo final se muestran en la Tabla 35. El promedio de AUC fue 84.57 % en Train, 84.29 % en Validación y 84.15 % en OOT (68.29 % de Gini). Gracias a la optimización bayesiana, se logró una métrica muy similar a LightGBM (68.9 % Gini OOT) y superior al XGBoost base (66.2 % Gini OOT). No existen grandes diferencias entre los tres conjuntos, lo que indica que el modelo generaliza correctamente. La Figura 48 muestra la curva ROC del Modelo.

Tabla 35.

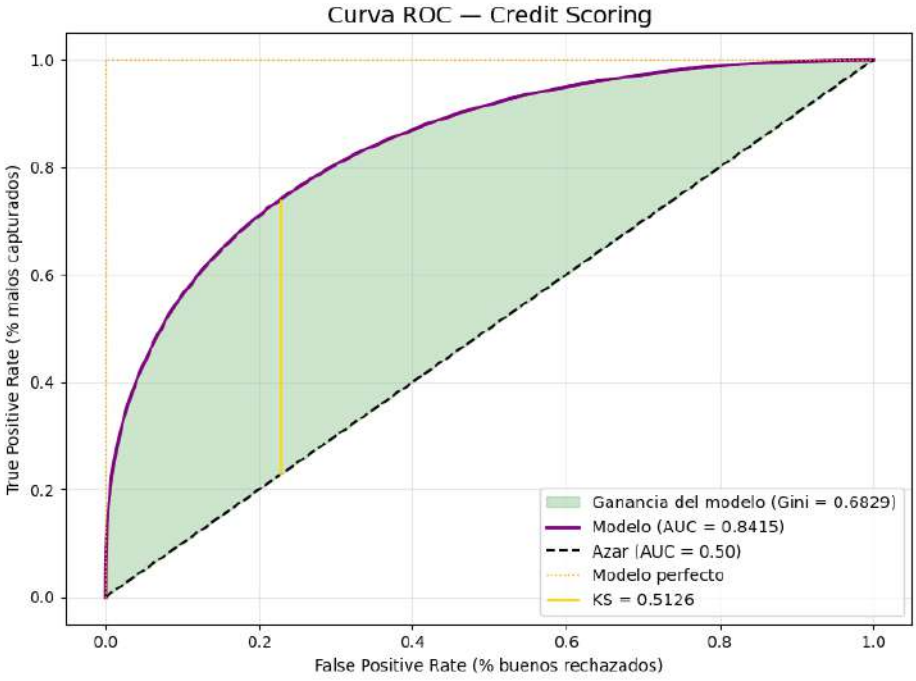
Resumen de Indicadores de Performance del Modelo XGBOOST Optimizado

Trial	Fold	Train AUC	Gini Train	Val AUC	Gini Val	Diff AUC	Diff Gini	AUC OOT	Gini OOT
#20	1	84.49 %	68.98 %	84.54 %	69.08 %	0.1 %	0.1 %	84.15 %	68.29 %
#20	2	84.60 %	69.20 %	84.18 %	68.36 %	-0.4 %	-0.8 %		
#20	3	84.55 %	69.10 %	84.36 %	68.72 %	-0.2 %	-0.4 %		
#20	4	84.57 %	69.14 %	84.22 %	68.44 %	-0.4 %	-0.7 %		
#20	5	84.62 %	69.24 %	84.13 %	68.26 %	-0.5 %	-1.0 %		
Promedio		84.57 %	69.13 %	84.29 %	68.6 %	-0.3 %	-0.6 %	-0.14 %	-0.28 %

Nota. Fuente: Elaboración propia

Figura 48.

Curva ROC del Modelo XGBOOST Optimizado



Nota. Fuente: Elaboración propia

El Gini permite conocer la capacidad predictiva, un Gini alto diferencia de forma fiable entre prestatarios con probabilidades de impago y los que no, garantizando decisiones técnicamente sólidas y éticas. Mittigová (2018) clasifica los rangos de AUC y Gini en el mundo del riesgo crediticio según la Tabla 36.

Tabla 36.

Clasificación del Gini según Escalas

Rango de AUC	Rango de Gini	Clasificación Benchmarking Banca
0 % - 60 %	0 % - 20 %	Muy Malo
60 % - 70 %	20 % - 40 %	Malo
70 % - 80 %	40 % - 60 %	Regular
80 % - 90 %	60 % - 80 %	Bueno
90 % - 100 %	80 % - 100 %	Muy Bueno

Nota. Fuente: Elaboración propia

El modelo final presenta un Gini de 68.29 % en OOT, situándose en la categoría "Bueno". Esto demuestra una alta potencia discriminatoria, capturando con precisión las señales de impago y reduciendo la probabilidad de error en la clasificación.

El estadístico Kolmogorov-Smirnov (KS) mide la máxima distancia entre las funciones de distribución acumulada de eventos y no-eventos, según la Ecuación 4. Un valor de KS ≥ 0.4 se considera aceptable en el mundo crediticio (Barreto, 2025).

Ecuación 4:

Estadístico de Kolmogorov-Smirnov (KS)

$$KS = \max(|F_1(x) - F_0(x)|) \quad (4)$$

Donde:

- $F_1(x)$ = Función de distribución acumulada de positivos

- $F_0(x)$ = Función de distribución acumulada de negativos

En la Tabla 37 se detalla la escala para el estadístico de KS. El modelo obtuvo un KS de 0.51 en OOT, lo que representa una "Buena discriminación". El punto de corte óptimo se determinó mediante el Índice de Youden, que coincide con la máxima separación del KS. La Figura 49 muestra que el punto de corte obtenido: 0.5084. Valores superiores a esta probabilidad serán considerados como default.

Tabla 37.

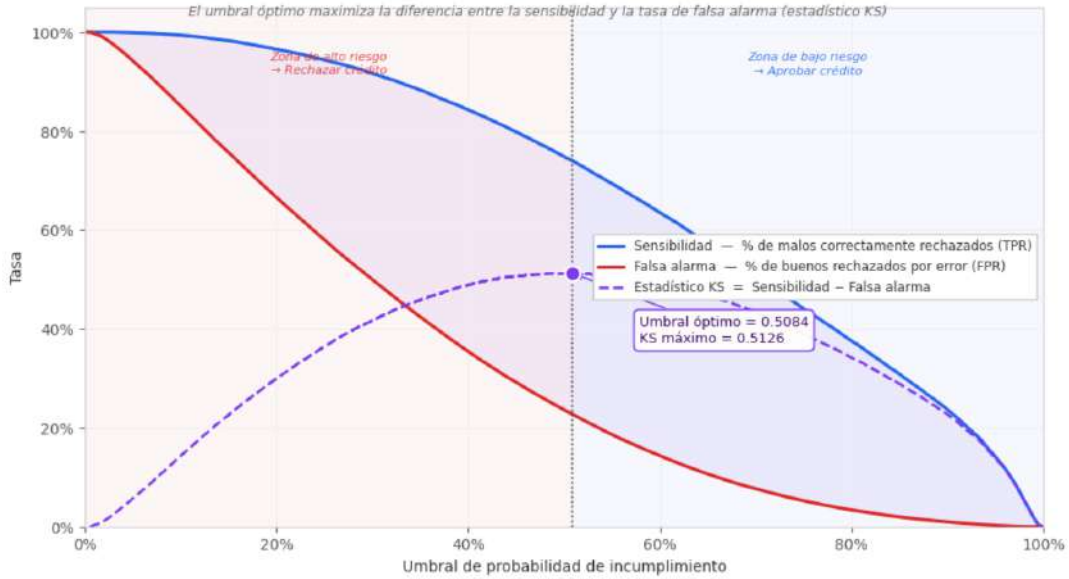
Clasificación del KS según Escalas

Valor de KS	Interpretación
KS < 0.2	Muy débil – Sin capacidad discriminante
0.2 ≤ KS < 0.3	Discriminación limitada
0.3 ≤ KS < 0.4	Discriminación aceptable
0.4 ≤ KS < 0.6	Buena discriminación
KS ≥ 0.6	Excelente discriminación

Nota. Fuente: Elaboración propia

Figura 49.

Selección de Punto de Corte Optimo – Método de Youden

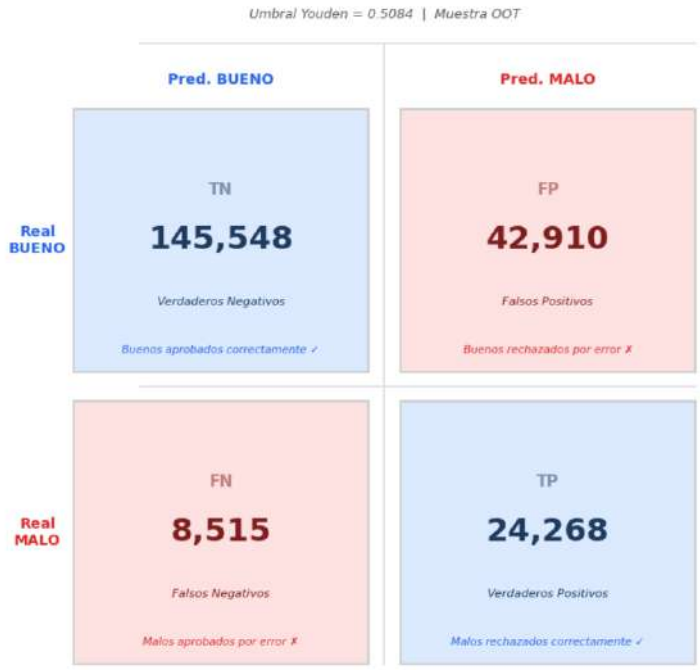


Nota. Fuente: Elaboración propia

La matriz de confusión y los indicadores derivados se muestran en la Figura 50 y la Tabla 38 respectivamente. Se obtuvo una accuracy del 76.8 %, sensibilidad del 74 % y especificidad del 77.2 %. En riesgo de crédito, la sensibilidad y especificidad son clave: la primera indica cuántos malos capturamos, la segunda cuántos buenos no rechazamos.

Figura 50.

Matriz de confusión – Vista por cuadrantes



Nota. Fuente: Elaboración propia

Tabla 38.

Metricas derivadas de la Matriz de Confusión

Indicador	Fórmula	Valor	Interpretación de negocio
Accuracy	(TP+TN) / Total	76.8 %	De cada 100 clientes, cuántos clasificó bien

Sensibilidad (Recall)	TP / (TP+FN)	74.0 %	De cada 100 malos reales, cuántos detectó
Especificidad	TN / (TN+FP)	77.2 %	De cada 100 buenos reales, cuántos aprobó
Precisión (VPP)	TP / (TP+FP)	36.1 %	De cada 100 rechazados, cuántos eran malos
VPN	TN / (TN+FN)	94.5 %	De cada 100 aprobados, cuántos eran buenos
F1 Score	2×Prec×Rec / (Prec+Rec)	48.6 %	Balance entre precisión y sensibilidad
FPR	FP / (FP+TN)	22.8 %	De cada 100 buenos, cuántos rechazó por error
FNR	FN / (FN+TP)	26.0 %	De cada 100 malos, cuántos dejó pasar por error

Nota. Fuente: Elaboración propia

4.1.3.6 Transformación a Score Crediticio (Scorecard)

Para hacer más interpretable el output del modelo (probabilidad de default), se aplicó una normalización. Metawa et al. (2025) indica que el score crediticio es una normalización en escala de 0 a 1000, se aplicó la Ecuación 5. Esta transformación entrega un valor más entendible para el negocio: valores cercanos a 1000 indican menor riesgo, y valores cercanos a 0 indican mayor riesgo.

Ecuación 5:

Transformación a Score Crediticio (Scorecard)

$$Score = (1 - p)x1000 \quad (5)$$

Donde:

- p = Probabilidad (Output del modelo)

4.1.4 Diseño e implementación del sistema de seguimiento de performance, estabilidad e interpretabilidad del modelo

Para realizar la integración del Modelo y la Plataforma analítica de seguimiento con las mejores prácticas de la Ingeniería de Software se desarrolló una especificación formal de requisitos. Las necesidades analíticas y operativas del negocio fueron capturadas mediante Historias de Usuario. Las cuales se tradujeron en Requisitos Funcionales y No Funcionales evaluados bajo la ISO 25010. Mediante el documento de Análisis de Software se recoge la necesidad de los usuarios tanto para el producto de la plataforma web como para lo planteado anteriormente en el desarrollo del modelo (Para más detalle véase el Anexo A – Documento de Análisis de Software). Este marco fue el punto de partida para guiar el diseño de la arquitectura en la nube, la interfaz de usuario y el agente basado en LLM que se exponen en las siguientes secciones.

4.1.4.1 Análisis de Interpretabilidad Global e Individual mediante Valores de Shapley (SHAP)

Para analizar la interacción e importancia de las variables, se calcularon los Shapley Values (SHAP) que es método que proporciona un marco para interpretar los resultados obtenidos, además de detectar dependencias no lineales entre variables. En la Tabla 39 se muestran los valores SHAP y el peso relativo de cada variable en el modelo final. Las cinco variables más importantes son: % de Uso TC último mes, Mínimos pasivos totales últimos 3 meses, Máximos días de atraso Riesgo Directo Personal Natural últimos 3 meses, Tendencia de Activos totales últimos 12 meses y Máximos días de atraso riesgo directo persona natural último año. Estas cinco concentran el 28 % del peso total de las 48 variables.

Tabla 39.

Top Variables mas Importantes de Modelo XGBOOST Final

Top Variable	Variable	Descripción	SHAP Medio	SHAP Relativo
1	pct_uso_tctot_u1m	% de Uso Total de Tarjeta de Crédito Último Mes	0.352438	8.9%
2	min_pas_tot_3m	Mínima Cantidad de Pasivos Total Últimos 3 Meses	0.250986	6.4%
3	max_datraso_rd_pn_u3m	Máximo Días de Atraso Riesgo Directo Persona Natural Últimos 3 Meses	0.177632	4.5%
4	beta_activos_total_12m	Tendencia de Activos Totales Últimos 12 Meses	0.154835	3.9%
5	datraso_rd_pn_u1m	Días de Atraso Riesgo Directo Persona Natural Último Mes	0.149137	3.8%
6	flag_dependiente	Indicador de Cliente Dependiente	0.148851	3.8%
7	max_datraso_rd_pn_u1a	Máximo Días de Atraso Riesgo Directo Persona Natural Último Año	0.1449	3.7%
8	prom_dec_12m	Promedio de Decrementos Último Año	0.127502	3.2%
9	ratio_pascur_pasavg_3m	Ratio Pasivos Fin de Mes entre Pasivos Promedio Últimos 3 Meses	0.118135	3.0%
10	min_pas_sueldo_12m	Mínimo Pasivo Cuenta Sueldo Últimos 12 Meses	0.111627	2.8%
11	saldo_pp_u1m	Saldo Préstamos Personales Último Mes	0.105691	2.7%
12	cuenta_decre_12m	Número de Decrementos Último Año	0.104652	2.7%
13	beta_entidades_dist_rd_lineatc_u1a	Tendencia de Entidades con Deuda en Tarjeta de Crédito Último Año	0.091527	2.3%
14	prm_ddatelc12m	Promedio de Deuda en Telecomunicaciones Último Año	0.087674	2.2%
15	nro_producto_6m	Número de Intenciones de Acceder al Producto Últimos 6 Meses	0.086636	2.2%
16	nro_pagotarjeta_1m	Número de Pagos de Tarjetas Último Mes	0.086473	2.2%
17	edad	Edad del Cliente	0.085445	2.2%
18	desv_saldo_hip_u3m	Desviación del Saldo Hipotecario Últimos 3 Meses	0.084963	2.2%
19	prom_ingreso_u12m	Promedio de Ingreso Último Año	0.081486	2.1%
20	lvl_edu	Nivel Educativo	0.078691	2.0%

21	max_decrem_12m	Máximo Decremento Últimos 12 Meses	0.07674	1.9%
22	prm_trxtot12m	Promedio de Transacciones Totales Últimos 12 Meses	0.074226	1.9%
23	beta_saldo_tc_tot_u1a	Tendencia Saldo Tarjeta de Crédito Total Último Año	0.070736	1.8%
24	linea_tc_tot_u1m	Línea Total Tarjeta de Crédito Último Mes	0.067997	1.7%
25	prm_telcorep12m	Promedio de Empresas que Reportan Deuda en Telecomunicaciones Último Año	0.067429	1.7%
26	cuenta_decre_12m_tc	Número de Decrementos en Tarjeta de Crédito Último Año	0.06394	1.6%
27	procedencia	Procedencia del Cliente	0.06129	1.6%
28	prm_trxdispefect12m	Promedio de Transacciones de Disposición de Efectivo Último Año	0.056827	1.4%
29	pct_saldo_rd_g1246_u1m	% de Saldo Riesgo Directo en Principales Entidades Último Mes	0.054989	1.4%
30	ciiu_agrup	Agrupación de Actividad Económica	0.052071	1.3%
31	beta_saldo_pp_bank_1a	Tendencia Saldo Interno Préstamos Personales Último Año	0.050672	1.3%
32	beta_cur_pas_total_12m	Tendencia Pasivo a Fin de Mes Últimos 12 Meses	0.050019	1.3%
33	beta_saldo_tc_tot_u6m	Tendencia Saldo Total Tarjeta de Crédito Últimos 6 Meses	0.046996	1.2%
34	prm_trxorggub12m	Promedio de Transacciones en Organizaciones Gubernamentales Último Año	0.046648	1.2%
35	ratio_act_pas_tot_12m	Ratio entre Activos y Pasivos Últimos 12 Meses	0.045075	1.1%
36	beta_saldo_hip_u1a	Tendencia Saldo Hipotecario Último Año	0.043343	1.1%
37	saldo_disp_ef_pn_u1m	Saldo Disposición de Efectivo Persona Natural Último Mes	0.043011	1.1%
38	max_decre_u03m_pp_bank	Máximo Decremento Interno Préstamos Personales Últimos 3 Meses	0.042544	1.1%
39	prm_trxrestaurant12m	Promedio de Transacciones en Restaurantes Últimos 12 Meses	0.039994	1.0%

40	segmento_rfm_fin	Segmento RFM Financiero	0.039089	1.0%
41	beta_pct_uso_tcbank_u1a	Tendencia % de Uso Interno Tarjeta de Crédito Último Año	0.038453	1.0%
42	beta_saldo_pp_u1a	Tendencia Saldo Pasivo Préstamos Personales Último Año	0.037283	0.9%
43	rto_decre_saldo_u3m_bank	Ratio Decremento entre Saldo Últimos 3 Meses Interno	0.032985	0.8%
44	segmento_ssff_tc	Segmento Sistema Financiero Tarjeta de Crédito en Función del Riesgo	0.030609	0.8%
45	promingr_xhog22	Promedio Gasto Mensual del Hogar entre 2016 y 2022	0.027368	0.7%
46	prm_trxhospedaj12m	Promedio de Transacciones en Hospedajes Último Año	0.023408	0.6%
47	desv_entidades_dist_rd_lineatc_u6m	Desviación de Entidades con Deuda en Tarjeta de Crédito Últimos 6 Meses	0.021446	0.5%
48	max_decre_u06m_pp_bank	Máximo Decremento Interno Préstamos Personales Últimos 6 Meses	0.008478	0.2%

Nota. Fuente: Elaboración propia

Una de las ventajas de SHAP es que permite interpretar y visualizar el impacto de cada variable en la predicción final. En la Figura 51 se aprecia el valor log-odds final, que es de 1.66 para un registro de la muestra OOT. Al aplicar la función sigmoide, se obtiene una probabilidad de 0.839 de pertenecer a la clase positiva. Se observa cómo la variable Porcentaje de Uso TC Total último mes empuja (castiga) la predicción en 0.99 log-odds, mientras que el Número de Decrementos últimos 12 meses la contrae (premia) en -0.7. La suma de todos estos log-odds más el valor base da el resultado final de 1.66.

Figura 51.

Force Plot para un registro del conjunto OOT

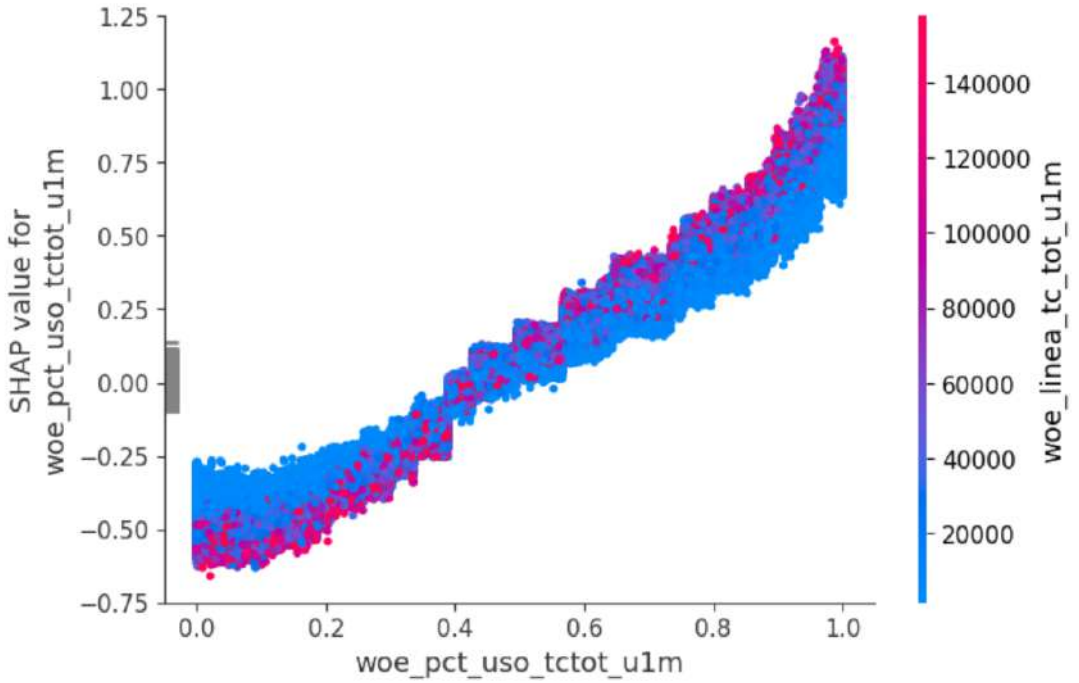


Nota. Fuente: Elaboración propia

Para una vista global del modelo, se generaron Gráficos de Dependencia Parcial. Estos muestran el cambio promedio en la predicción para todo el conjunto de datos. En el Gráfico 52 se aprecia la relación directa entre la variable top 1 y el target, es decir a medida que aumenta el valor de esta variable, la probabilidad de pertenecer a la clase 1 también aumenta.

Figura 52.

Gráfico de Dependencial Parcial Variable Top 1- % Uso TC último mes.

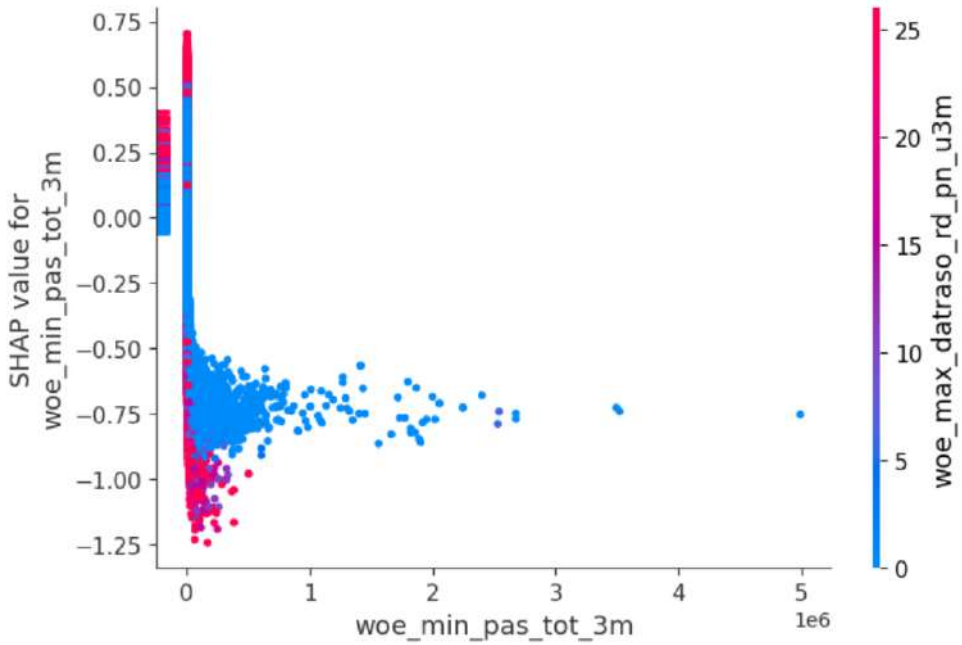


Nota. Fuente: Elaboración propia

En la Figura 53 se observa que la relación con el target es inversa para la variable top 2. Mientras crece el valor de esta variable, la probabilidad de ser clase 1 disminuye.

Figura 53.

Gráfico de Dependencial Parcial Variable Top 2 – Minimo pasivos u3m

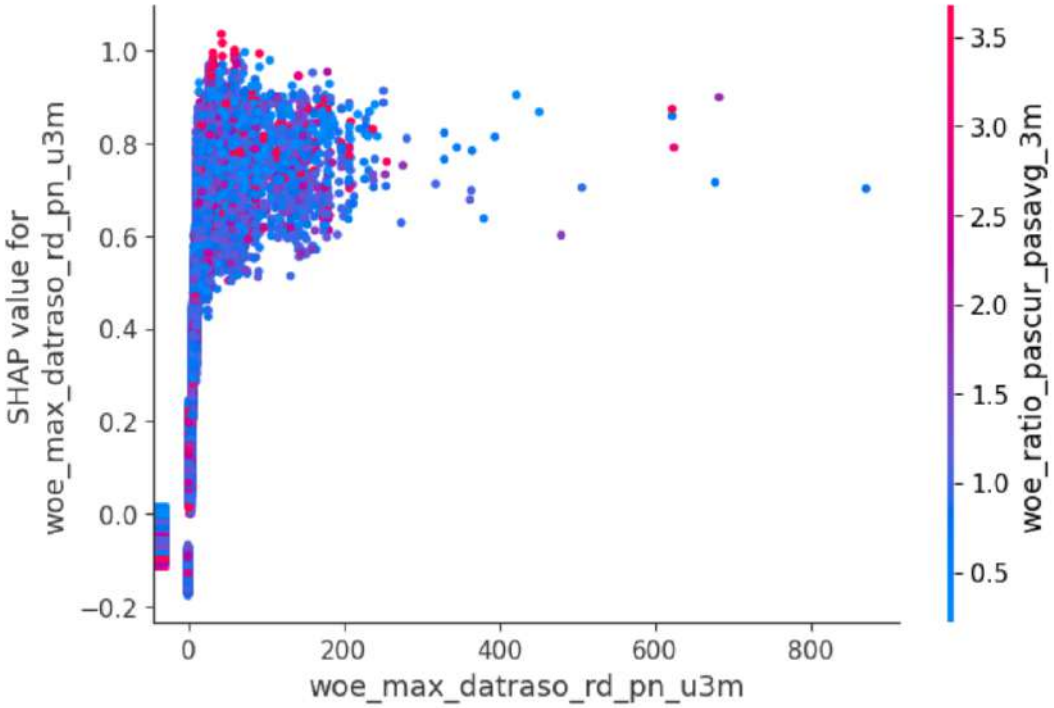


Nota. Fuente: Elaboración propia

La Figura 54, explica que la variable top 3 muestra una relación directa: al aumentar el valor de la variable, crece la probabilidad de ser clase 1.

Figura 54.

Gráfico de Dependencial Parcial Variable Top 3 – Maximo dias de atraso riesgo directo persona natural u3m

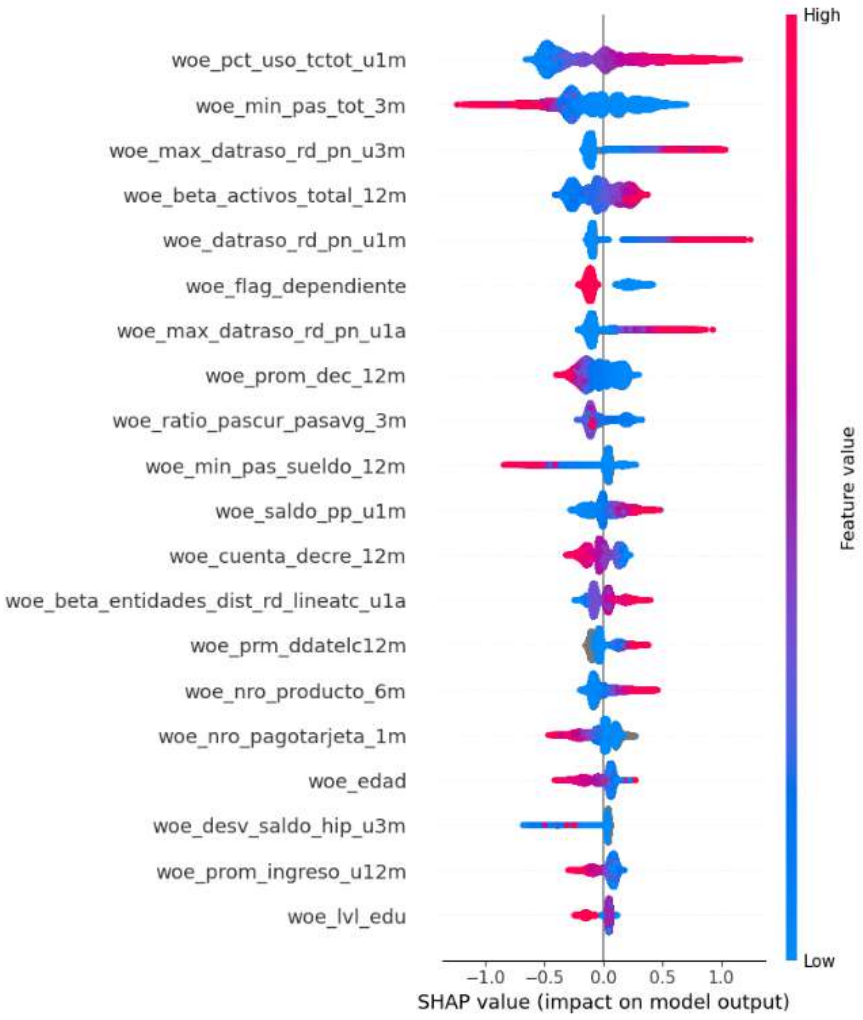


Nota. Fuente: Elaboración propia

Finalmente, el resumen del TOP 20 de variables se observa en la Figura 55 evidencia cómo cada variable contribuye a aumentar o disminuir la probabilidad de default. Variables como d_atraso_rd_pn_u1m y beta_entidades-distr_rd_linetc incrementan el riesgo, mientras que flag_dependiente y prom_dec_12m lo reducen.

Figura 55.

Gráfico de Resumen de tendencia SHAP para Top 20 Variables



Nota. Fuente: Elaboración propia

4.1.4.2 Arquitectura en Cloud para el Cálculo de Indicadores de Monitoreo

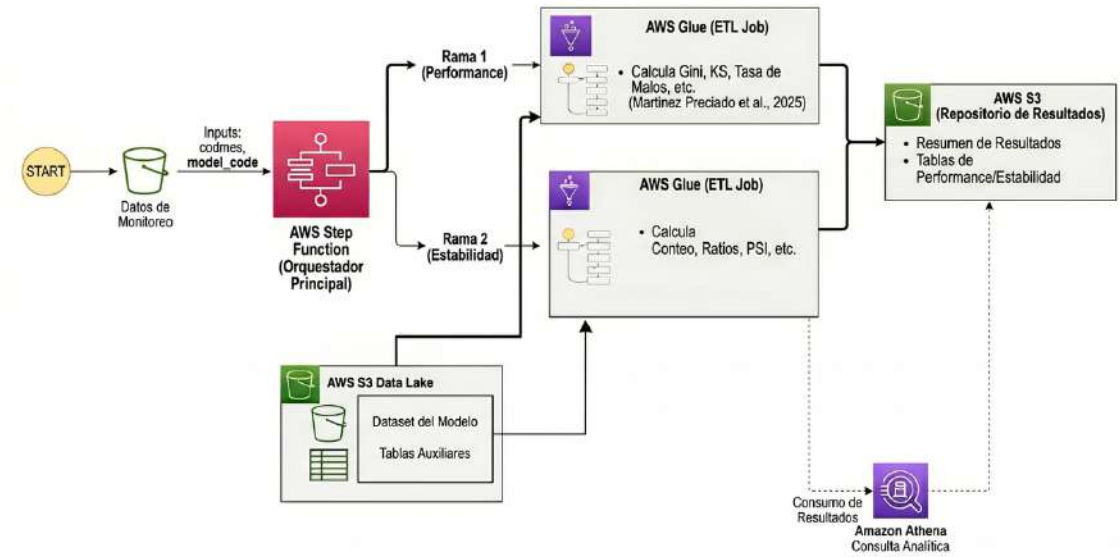
El monitoreo sistemático de modelos predictivos es crítico en la gestión del riesgo analítico en el sector regulado. El seguimiento periódico es una exigencia técnico-regulatoria indispensable para mitigar el riesgo latente de los algoritmos (Martinez Preciado et al., 2024). Este riesgo se manifiesta en posibles pérdidas financieras o decisiones operativas erróneas debido a fallas en el desarrollo o uso de modelos cuantitativos (Para más detalle véase el Anexo B – Documento de Arquitectura de Software). La importancia del monitoreo se sustenta en cuatro pilares:

- Verificación del desempeño real del algoritmo frente a las dinámicas cambiantes del mercado.
- Mitigación de riesgos operativos originados por herramientas desactualizadas.
- Detección temprana de alertas ante anomalías estadísticas.
- Validación continua de que la población objetivo no haya sufrido variaciones estructurales respecto al diseño muestral original.

Para desplegar este control de manera automatizada y escalable, se diseñó una arquitectura en Amazon Web Services (AWS) usando Step Functions como orquestador del flujo de cálculo de métricas de monitoreo. AWS Step Functions permite coordinar múltiples servicios en flujos de trabajo desacoplados y resilientes. La Figura 56 muestra el detalle de la arquitectura.

Figura 56.

Diagrama de Arquitectura Serveless para la orquestación y cálculo de indicadores de Monitoreo



Nota. Fuente: Elaboración propia

El flujo de procesamiento e ingesta de datos sigue esta secuencia técnica:

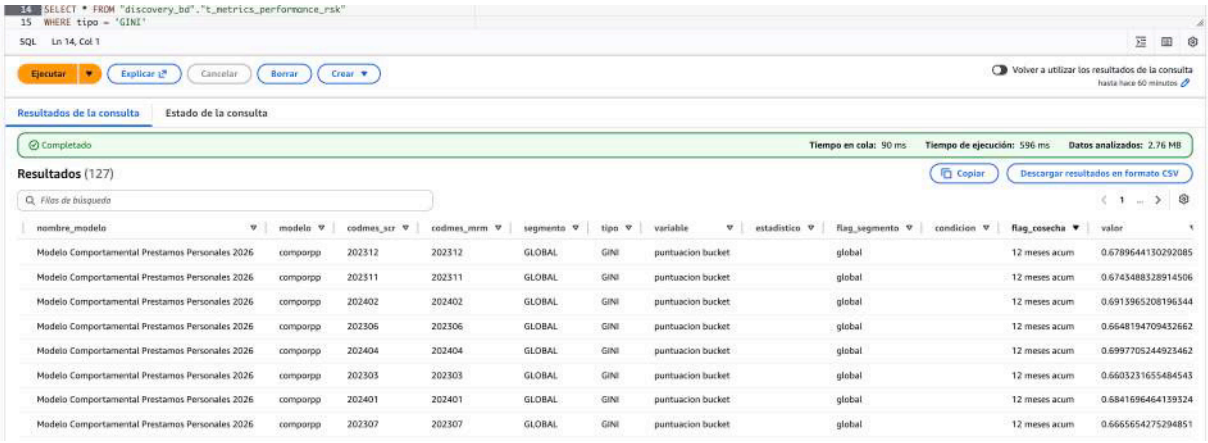
- Desencadenamiento y Parametrización: El pipeline es gobernado por una función de estado (Step Function) que recibe como argumentos el mes de proceso (codmes) y el código identificador del modelo.
- Procesamiento Paralelo (AWS Glue): A partir de los parámetros de entrada, la función coordina la ejecución simultánea de dos componentes ETL:
    - Módulo de Performance: Consume el dataset histórico y calcula métricas predictivas mensuales y acumuladas: Gini, KS, tasa de malos (bad rate) y matriz de confusión. Para Gini y tasa de malos, el cálculo se realiza tanto a nivel de score como de variables del modelo.
    - Módulo de Estabilidad: Evalúa mensual y acumuladamente los cambios en la estructura de los datos de entrada: conteo de registros, ratios de distribución e Índice de Estabilidad Poblacional (PSI).
- Estrategia de Discretización Dinámica (Bucketing): Para construir los intervalos necesarios en el cálculo del PSI y la tasa de malos, el componente consume tablas auxiliares con los puntos de corte (splits) del modelo. La lógica aplica:
    - Cortes Óptimos: Se priorizan los límites precalculados con la librería Optimal Binning.
    - Variables Discretas: Se respeta la cantidad de grupos iniciales (ej. Segmento RFM, Procedencia).
    - Quintiles: Si no hay cortes óptimos, se generan cinco grupos de igual tamaño.
    - Deciles: El procesamiento del score final se ejecuta con partición de diez grupos.

- Persistencia Estructurada: Los resultados procesados por ambos módulos se almacenan en un bucket S3, para luego ser consultados mediante Athena con SQL.

En la Figura 57 se muestra un ejemplo de los resultados de performance para el indicador GINI.

Figura 57.

Tabla de Resultados de Performance consultada mediante Athena para indicador Gini



nombre_modelo	modelo	codmes_scr	codmes_mrm	segmento	tipo	variable	estadistico	flag_segmento	condicion	flag_cosecha	valor
Modelo Comportamental Prestamos Personales 2026	comporpp	202312	202312	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6789644130292085
Modelo Comportamental Prestamos Personales 2026	comporpp	202311	202311	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6743488328914506
Modelo Comportamental Prestamos Personales 2026	comporpp	202402	202402	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6913965208196344
Modelo Comportamental Prestamos Personales 2026	comporpp	202306	202306	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6648194709432662
Modelo Comportamental Prestamos Personales 2026	comporpp	202404	202404	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6997705244923462
Modelo Comportamental Prestamos Personales 2026	comporpp	202303	202303	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6603231655484543
Modelo Comportamental Prestamos Personales 2026	comporpp	202401	202401	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6841696464139324
Modelo Comportamental Prestamos Personales 2026	comporpp	202307	202307	GLOBAL	GINI	puntuacion bucket		global		12 meses acum	0.6665654275294851

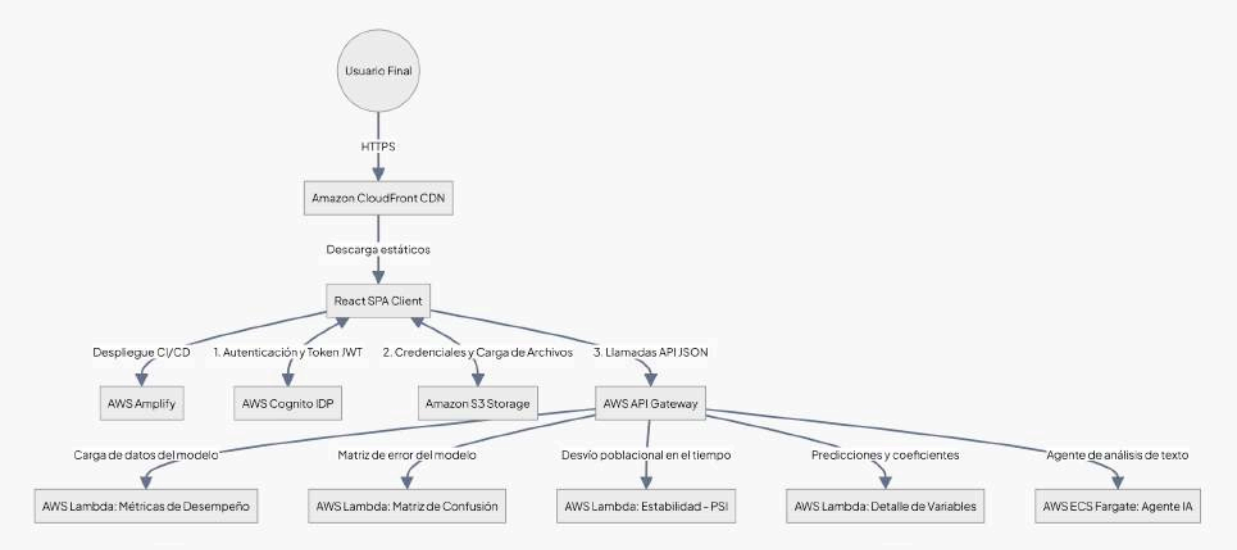
Nota. Fuente: Elaboración propia

4.1.4.3 Despliegue de Plataforma Web de Monitoreo: Model Risk Management Suite

Para consumir y analizar los resultados del modelo, se desplegó una plataforma web con arquitectura desacoplada bajo el paradigma Single Page Application (SPA) en el cliente. Se comunica mediante APIs con una infraestructura en la nube bajo el modelo Serverless Computing (FaaS) en AWS. Esta estrategia garantiza tolerancia a fallos nativa, escalabilidad automática y costos operativos proporcionales a la demanda. El diagrama técnico de la Figura 58 detalla el flujo de datos desde el navegador web a través de los componentes de la plataforma (Para más detalle véase el Anexo B – Documento de Arquitectura de Software).

Figura 58.

Diagrama General de la Plataforma Web



Nota. Fuente: Elaboración propia

Para garantizar que el uso del usuario final sea eficiente y vaya de acorde a las necesidades planteadas, el desarrollo del frontend se guio por un diseño centrado en el usuario. El prototipo visual, la disposición de los tableros de reportera (performance y estabilidad) y la interfaz conversacional del Agente MRM fueron plasmados en wireframes que responden a las Historias de Usuario previamente definidas. (Para más detalle véase Anexo C - Documento de Diseño de Interfaz y Experiencia de Usuario UX/UI).

4.1.4.3.1 Flujo de Operaciones

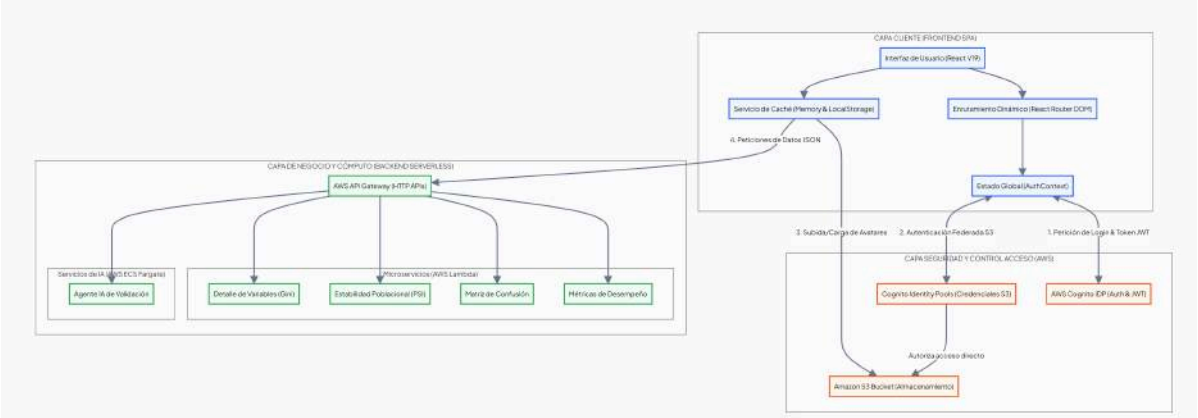
- Autenticación Federada: El cliente ingresa credenciales protegidas con AWS Cognito. El servicio verifica la cuenta y responde con un JSON Web Token (JWT) que contiene permisos y rol del analista.
- Acceso Seguro a Recursos (S3): Para cargar recursos (fotos de perfil), la app solicita credenciales temporales a través de Cognito Identity Pool, permitiendo interactuar directamente con S3 sin exponer llaves fijas.

- Consumo e Intermediación de APIs: Las solicitudes REST del cliente viajan firmadas con el token JWT. AWS API Gateway intercepta la llamada, valida el token y enruta hacia las Lambdas lógicas o hacia el Agente de IA.

  La Figura 59 muestra el diagrama general de arquitectura por capas.

Figura 59.

Diagrama de Arquitectura separado por Capas.



Nota. Fuente: Elaboración propia

4.1.4.3.2 Arquitectura Interna del Frontend (SPA)

La aplicación del lado del cliente se organiza en módulos con responsabilidades bien delimitadas:

- Enrutamiento y Control de Acceso Dinámico: Usa React Router DOM. La seguridad se resuelve con un componente ProtectedRoute que verifica autenticación y perfil completo en Cognito (custom: role, custom: department). Si falta algún atributo, redirige a /settings.
- Patrón Singleton de Caché en Memoria y Persistencia: Para evitar múltiples peticiones redundantes a las APIs, la capa de servicios implementa:
- Caché en Memoria (Singleton): Almacena respuestas JSON en variables globales.

- Caché en LocalStorage: Almacena datos serializados con clave de tiempo para sobrevivir a recargas.
- Control de Concurrencia: Si múltiples componentes invocan la misma consulta, retorna la misma promesa, evitando tráfico duplicado.
- Sincronización Optimizada de Estado: La comunicación entre AWS Cognito y la interfaz usa hooks de efecto optimizados. El estado solo se actualiza si los valores de los atributos han cambiado.

4.1.4.3.3 Arquitectura del Backend Serverless

El backend está completamente desacoplado de la visualización y se basa en servicios gestionados:

- AWS API Gateway (HTTP APIs): Enrutamiento de baja latencia, SSL, rate limiting y CORS predefinido.
- AWS Lambda (Compute Units): Funciones Python aisladas en micro-contenedores bajo demanda. Cada microservicio responde a un único tipo de datos lógicos (principio de responsabilidad única).
- AWS ECS & Fargate (Contenedor de IA): Para tareas continuas y computacionalmente intensivas. Hospeda la lógica del Agente IA mediante contenedor Docker.

4.1.4.4 Agente Model Risk Management (MRM) basado en Modelos de Lenguaje (LLM) y LangChain

Dentro de la plataforma web, se desplegó una sección especial denominada MRM - Agent. Este agente fue creado bajo el framework LangChain. K. S. Wang et al. (2025) indican que LangChain permite crear agentes con capacidad integral mediante recuperación, búsqueda

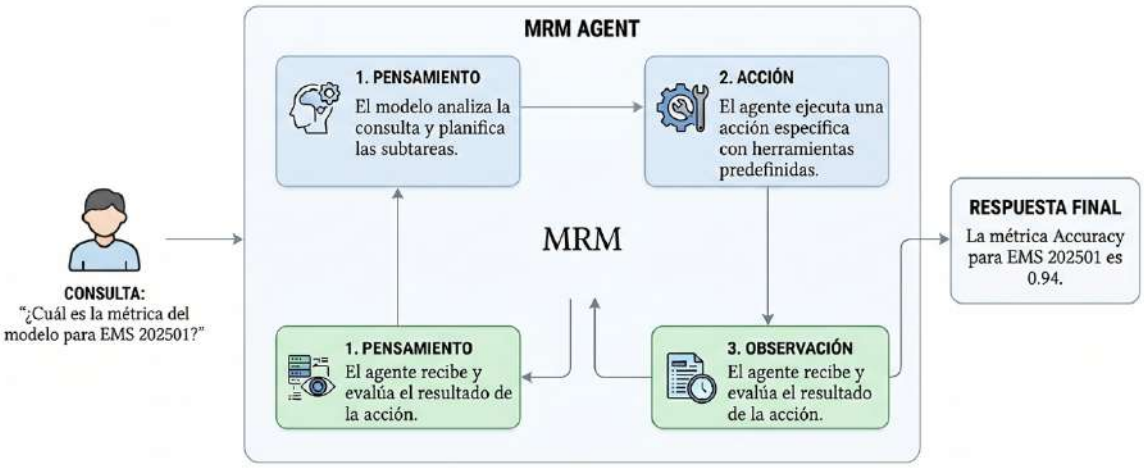
e interpretación. LangChain ofrece una plataforma unificada para trabajar con varios LLMs en Python.

4.1.4.4.1 Arquitectura del Agente: Integración del Patrón ReAct y LangChain

La arquitectura del agente sigue el patrón ReAct (Razonamiento, Actuación y Observación). Este patrón combina el razonamiento del modelo con herramientas externas. Yao et al. (2022) explican que los agentes basados en ReAct realizan un ciclo iterativo que se repite hasta cumplir el objetivo. La Figura 60 muestra el concepto de agente ReAct a alto nivel.

Figura 60.

Ejemplo de Agente Basado en patrón React



Nota. Fuente: Elaboración propia

4.1.4.4.2 Tools del Agente: Despliegue de herramientas y acceso a datos del proyecto

Las tools son el mecanismo que permite al agente ejecutar acciones sobre el entorno externo y funcionan como interfaces que conectan el LLM con fuentes de datos, APIs y sistemas externos. Para que la interacción sea efectiva, el LLM recibe una descripción clara de cada herramienta: función, parámetros de entrada y formato de salida. Así, el modelo puede seleccionar autónomamente la herramienta más adecuada según el contexto. En la Tabla 40 se

puede observar que se desplegaron 5 herramientas orientadas a cubrir las necesidades de acceso a información sobre el Model Risk Management.

Tabla 40.

Tools del Agente Model Risk Management

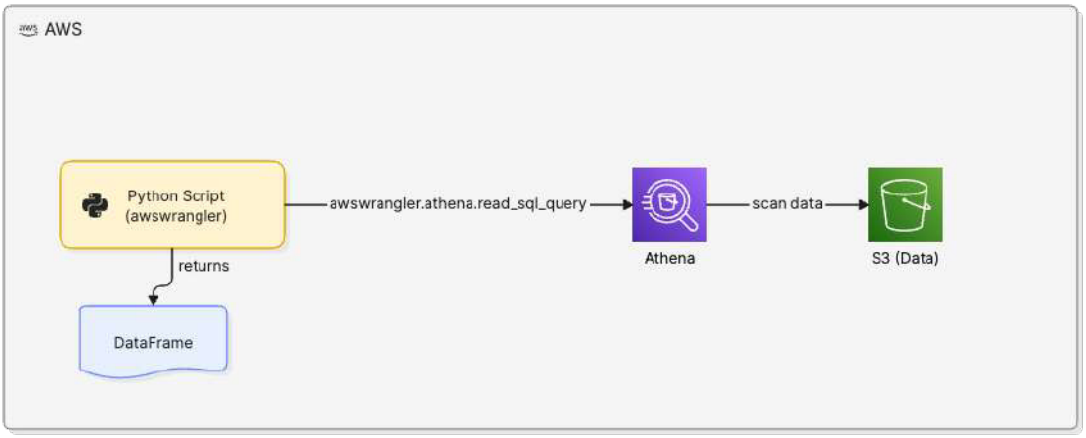
Herramienta	Descripción	Parámetros de entrada	Salida esperada
consultar_metricas_estabilidad	Consulta métricas de estabilidad de variables (PSI, count, ratio) para un mes específico.	codmes (YYYYMM), variable (opcional), metricas_solicitadas (psi/count/ratio)	Diccionario con métricas por variable y tramo, incluyendo PSI, distribución y conteos.
consultar_metricas_performance	Consulta métricas de performance del modelo (Gini, KS, matriz de confusión, ratio, TM) para un mes específico.	codmes_mrm (YYYYMM), variable (opcional), metricas_solicitadas (gini/ks/ratio/tm/matriz_confusion), maduracion_cosecha	Diccionario con indicadores macro y métricas por tramo.
consultar_maestro_cliente	Consulta el perfil general, saldos, productos y estado de gestionabilidad de un cliente.	cuc (10 dígitos), codmes (YYYYMM)	Diccionario con perfil demográfico, alertas de riesgo, vector de productos activos y estado de gestionabilidad.
consultar_valores_variables_cliente	Consulta el score y el set completo de variables explicativas de un cliente para uno o varios meses.	cuc (10 dígitos), codmes (YYYYMM o lista)	Diccionario con score y valores de variables por mes consultado.
consultar_valores_shap_cliente	Consulta la explicabilidad del modelo (valores SHAP en Log-Odds y puntos score) de un cliente autorizado.	cuc (10 dígitos), codmes (YYYYMM o lista)	Diccionario con aportes SHAP en Log-Odds y puntos de score por mes consultado.

Nota. Fuente: Elaboración propia

Estas herramientas consumen scripts de Python probados en SageMaker que recuperan datos de diferentes tablas de Athena. Por ejemplo, si un usuario usa consultar_metricas_estabilidad, esta llama a la tabla metrics_stability y devuelve el dataframe para su consumo. La Figura 61 muestra el flujo de estas herramientas.

Figura 61.

Diagrama de Alto Nivel de Funcion de Herramientas de Recuperación



Nota. Fuente: Elaboración propia

4.1.4.4.3 Mecanismo de Recuperación de Información y Generación Aumentada (RAG)

RAG (Retrieval-Augmented Generation) combina LLMs con sistemas de recuperación de información externa, permitiendo consultar fuentes de datos dinámicas y de dominio específico (Liang et al., 2025). En este proyecto, el paradigma se implementa mediante una herramienta adicional. El LLM no responde desde su conocimiento interno, sino que activa herramientas para recuperar información actualizada desde una base de datos vectorial, fundamentando sus respuestas en datos reales, esto mitiga las alucinaciones, supera la

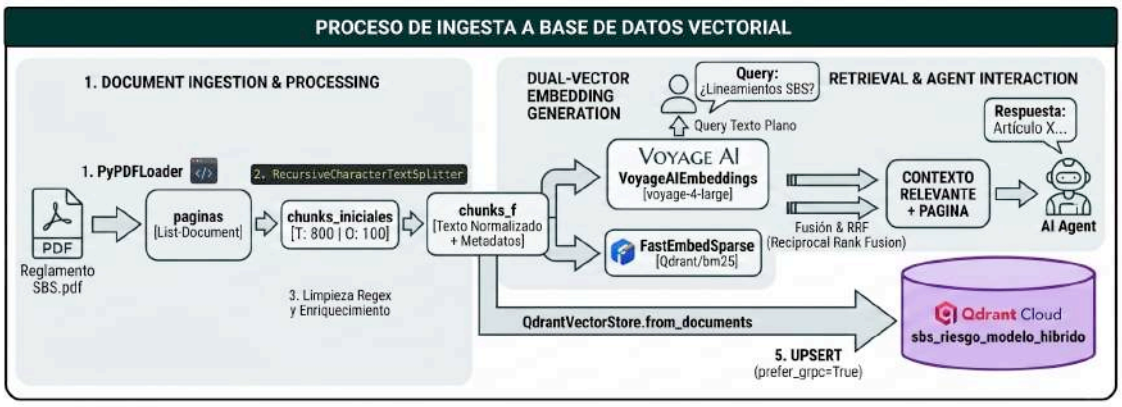
obsolescencia y evita reentrenamientos costosos. La implementación de RAG comprende dos fases:

- Fase de ingesta: Los datos de PDFs (manuales, reglamentos, documentos metodológicos) se procesan y estructuran para su posterior recuperación. Se usó el Loader de PDF para convertir a texto plano, luego se aplicó casteo a documents, chunkenizacion y enriquecimiento con metadatos. Se generaron dos embeddings: uno para búsqueda semántica con Voyage 4 large y otro para búsqueda textual con BM25. Finalmente, se almacenaron en una colección de QDRANT.
- Fase de recuperación: Ante una consulta, el LLM selecciona la herramienta adecuada, ejecuta la consulta en Athena, obtiene resultados y los integra en la respuesta final usando la técnica RRF (búsqueda híbrida).

La Figura 62 muestra el proceso de ingesta a la base de datos vectorial y la Figura 63 se puede apreciar el proceso de Retrieval o también llamado recuperación usando la técnica de RRF que permite aprovechar una búsqueda hibrida y obtener mejores resultados.

Figura 62.

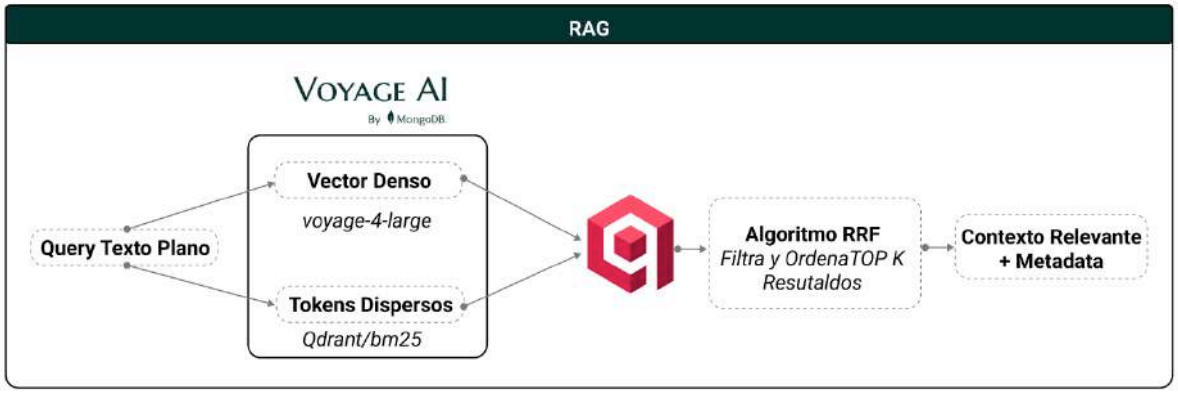
Proceso de Ingesta de Datos a Base de Datos Vectorial



Nota. Fuente: Elaboración propia

Figura 63.

Proceso de Recuperación de Datos para RAG



Nota. Fuente: Elaboración propia

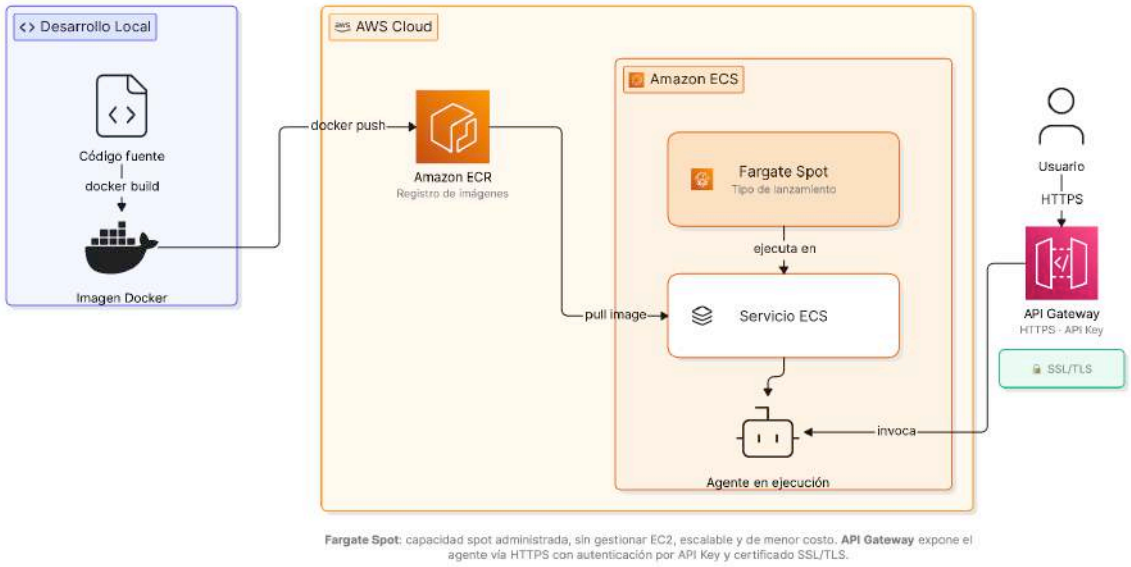
4.1.4.4.4 Despliegue del Agente como Microservicio

El agente se desplegó en producción contenerizado el código con Docker y orquestándolo en AWS Elastic Container Service (ECS). La imagen Docker se almacenó en Amazon Elastic Container Registry (ECR) y se desplegó como servicio en ECS con Fargate Spot, que permite ejecutar tareas en capacidad spot a menor costo. Esto garantiza un entorno administrado y escalable sin gestionar instancias EC2.

Para exponer el agente de manera segura a través de HTTPS, se implementó una capa de API Gateway que recibe peticiones del usuario y las redirige al servicio ECS Fargate. La API Gateway gestiona tráfico, autenticación mediante API Keys y proporciona un endpoint con certificado SSL/TLS, garantizando comunicación encriptada y segura. La Figura 64 muestra el proceso del despliegue del agente.

Figura 64.

Despliegue del Agente en Amazon Web Services (AWS)

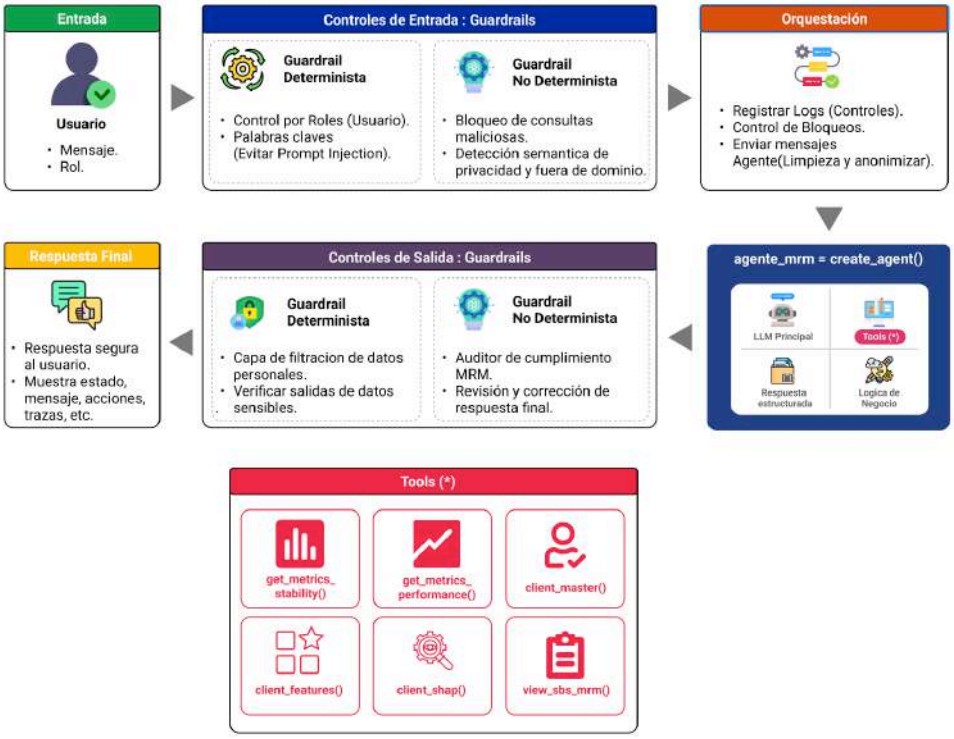


Nota. Fuente: Elaboración propia

Finalmente, la Figura 65 muestra el flujo completo del agente desplegado. Dispone de 6 herramientas: las primeras 5 provienen de consultas a bases de datos, la última del proceso RAG. Se implementaron Guardrails para evitar filtración de datos sensibles o desviaciones del dominio MRM.

Figura 65.

Flujo de Interacción Completo del Agente y Estructura General



Nota. Fuente: Elaboración propia

4.2. Validación del Sistema

La validación del sistema no fue un paso secundario, fue una fase deliberada para asegurar que el producto completo modelo, plataforma y agentes realmente aportara valor a los usuarios finales. Para cubrir todas las aristas, se dividió la validación en 3 frentes complementarios.

- Para el primero se definió y ejecutó un conjunto de pruebas funcionales y no funcionales orientadas a verificar el correcto funcionamiento de los principales casos de uso y atributos de calidad del sistema. Las pruebas funcionales incluyen casos de uso orientados a las principales historias de usuario. Asimismo, se consideraron pruebas no funcionales relacionadas con calidad web mediante PageSpeed Insights y rendimiento bajo carga/estrés concurrente con el software JMETER (Para más detalle véase el Anexo D – Documento de Pruebas de Software).

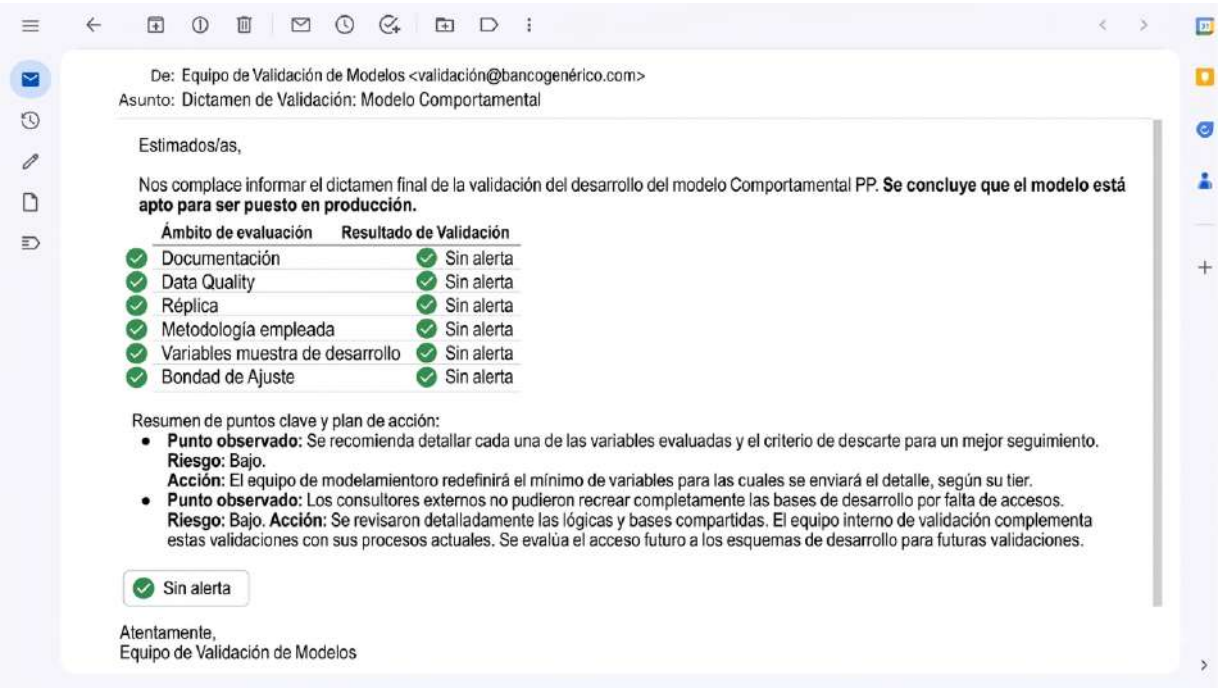
- El segundo, la validación del artefacto del modelo de probabilidad de impago, estuvo a cargo de una consultora internacional especializada en validaciones de modelos de riesgo. Su revisión cubrió la replicabilidad del código, la solidez de la metodología y la bondad de ajuste de las métricas.
- El tercero, la validación de la plataforma web Model Risk Management Suite, se abordó desde la óptica de los usuarios. Se diseñó una evaluación centrada en cuatro criterios: rendimiento, funcionalidad, seguridad y satisfacción general. Para ello, se recurrió a un comité de expertos del equipo de seguimiento y validación, quienes proporcionaron una valoración objetiva sobre cada uno de los criterios establecidos.

4.2.1 Validación del Artefacto del Modelo de Probabilidad de Impago

Para cumplir con la validación metodológica del artefacto, se realizó una revisión exhaustiva de todo el desarrollo del modelo de probabilidad de impago. Un consultor externo llevó a cabo esta validación, evaluando el modelo en tres frentes principales (Para más detalle véase el Anexo E – Validación del artefacto Modelo probabilidad de impago). La revisión metodológica dio como resultados lo mostrado en la Figura 66. El modelo obtuvo el visto bueno para su puesta en producción, superando todos los frentes de validación necesarios. La consultora también emitió algunas recomendaciones para futuros desarrollos.

Figura 66.

Conforme de la Validación del Artefacto del Modelo de Probabilidad de Impago



De: Equipo de Validación de Modelos <validación@bancogenérico.com>
Asunto: Dictamen de Validación: Modelo Comportamental

Estimados/as,

Nos complace informar el dictamen final de la validación del desarrollo del modelo Comportamental PP. **Se concluye que el modelo está apto para ser puesto en producción.**

Ámbito de evaluación	Resultado de Validación
Documentación	Sin alerta
Data Quality	Sin alerta
Réplica	Sin alerta
Metodología empleada	Sin alerta
Variables muestra de desarrollo	Sin alerta
Bondad de Ajuste	Sin alerta

Resumen de puntos clave y plan de acción:

- **Punto observado:** Se recomienda detallar cada una de las variables evaluadas y el criterio de descarte para un mejor seguimiento. **Riesgo:** Bajo. **Acción:** El equipo de modelamientoro redefinirá el mínimo de variables para las cuales se enviará el detalle, según su tier.
- **Punto observado:** Los consultores externos no pudieron recrear completamente las bases de desarrollo por falta de accesos. **Riesgo:** Bajo. **Acción:** Se revisaron detalladamente las lógicas y bases compartidas. El equipo interno de validación complementa estas validaciones con sus procesos actuales. Se evalùa el acceso futuro a los esquemas de desarrollo para futuras validaciones.

Sin alerta

Atentamente,
Equipo de Validación de Modelos

Nota. Fuente: Correo Enviado por la Consultara Internacional

4.2.2 Validación de la Plataforma WEB – Model Risk Management Suite

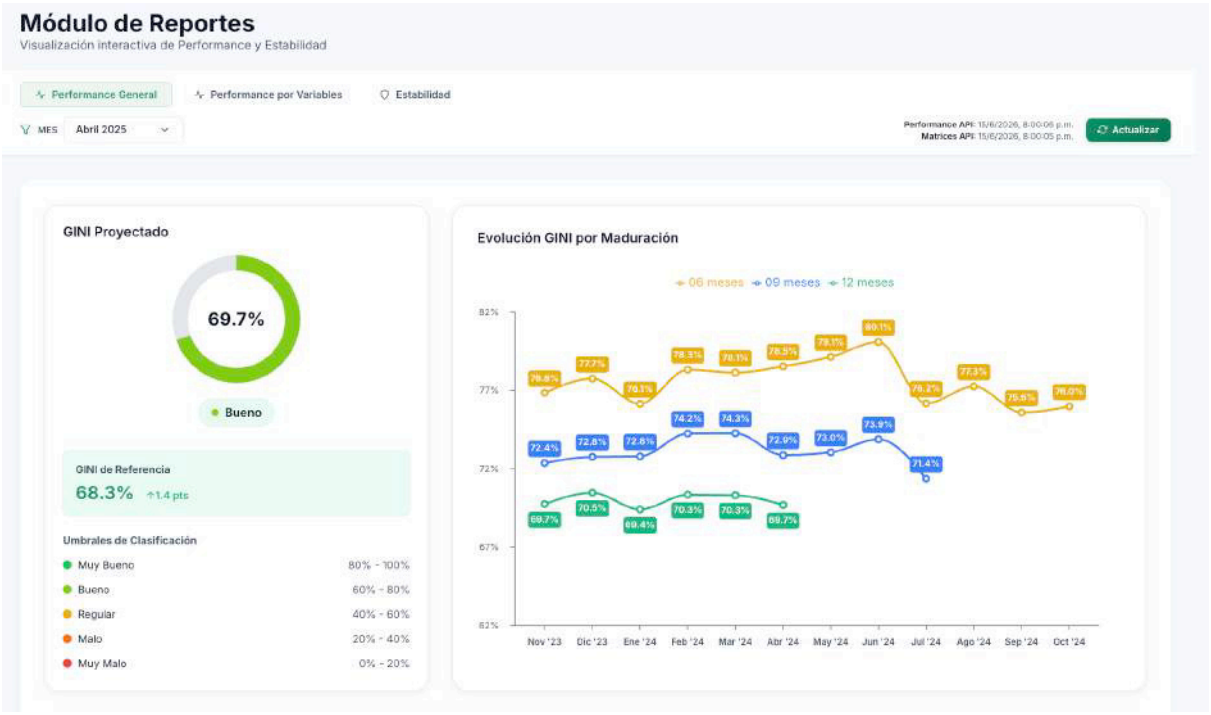
Para validar la solución completa, se puso la plataforma desplegada a disposición de usuarios internos del equipo de seguimiento y validación, los cuales interactuaron entre las siguientes principales vistas.

4.2.2.1 Monitoreo del Modelo Performance y Estabilidad

En la parte de seguimiento o monitoreo, se evalúa el rendimiento evolutivo del modelo en los meses posteriores a su desarrollo. En la Figura 67 se aprecia que el modelo alcanza casi 70 puntos de Gini en su último mes de monitoreo (abril de 2025), lo que lo coloca en la escala de "Bueno". Esta métrica está incluso 1.4 puntos por encima de su periodo OOT, lo que refleja un modelo robusto a pesar del paso de los meses. Para maduraciones tempranas (6 y 9 meses) se observa una tendencia similar.

Figura 67.

Vista de Performance General disponible en la plataforma Model Risk Management Suite

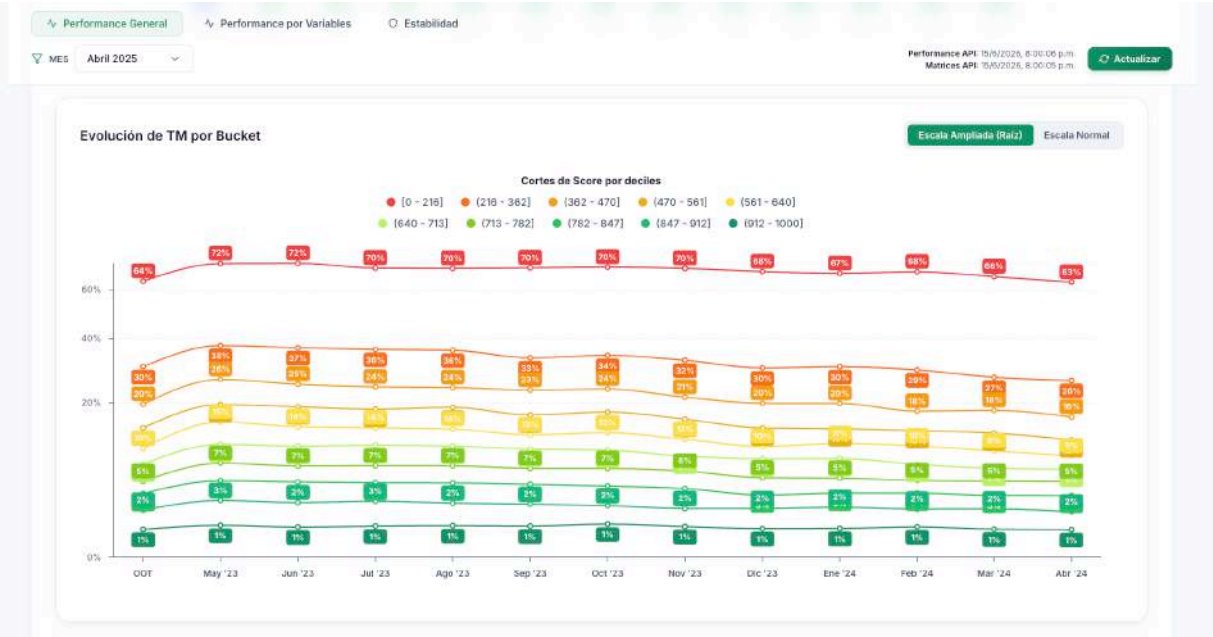


Nota. Fuente: Elaboración propia

También se validó la tasa de discriminación del score para los deciles de distribución. En la Figura 68 se aprecia que el modelo mantiene el ordenamiento para cada uno de los cortes, lo que concuerda con lo mostrado en la métrica de Gini.

Figura 68.

Vista de TM por decil disponible en la plataforma Model Risk Management Suite

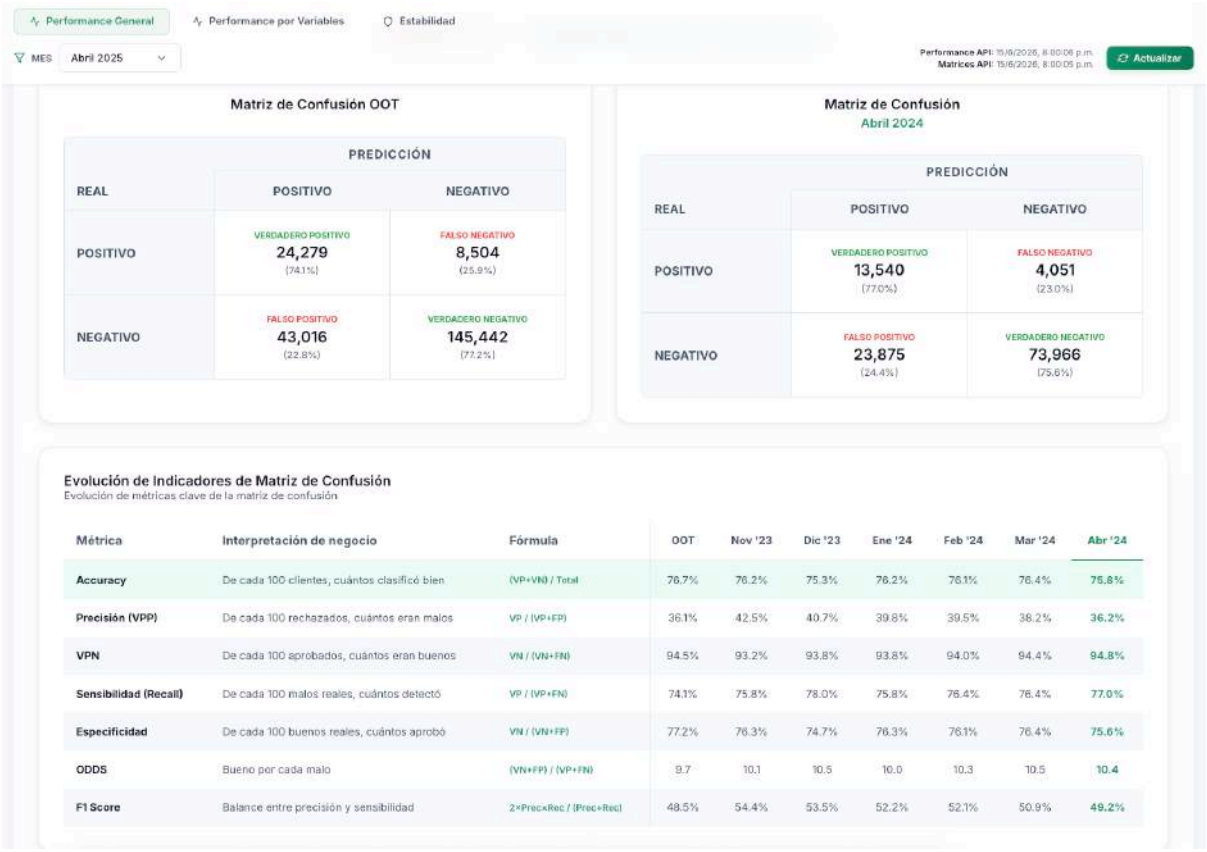


Nota. Fuente: Elaboración propia

Complementariamente, se dispone de vistas del evolutivo de la matriz de confusión (MC), el Gini por variables y la tasa de malos para cada variable. En todas estas vistas se observa que el modelo mantiene su poder predictivo en términos generales, reflejando un modelo integral. La Figura 69 muestra el evolutivo de las métricas de la matriz de confusión, las cuales se mantienen estables. La Figura 70 presenta una vista de la tasa de malos (tm) para la variable más importante del modelo, la cual conserva el ordenamiento para cada uno de sus cortes.

Figura 69.

Vista de Indicadores de MC disponible en la plataforma Model Risk Management Suite



Nota. Fuente: Elaboración propia

Figura 70.

Comparativo de Tasa de Malos para variable TOP 1 disponible en la plataforma Model Risk Management Suite

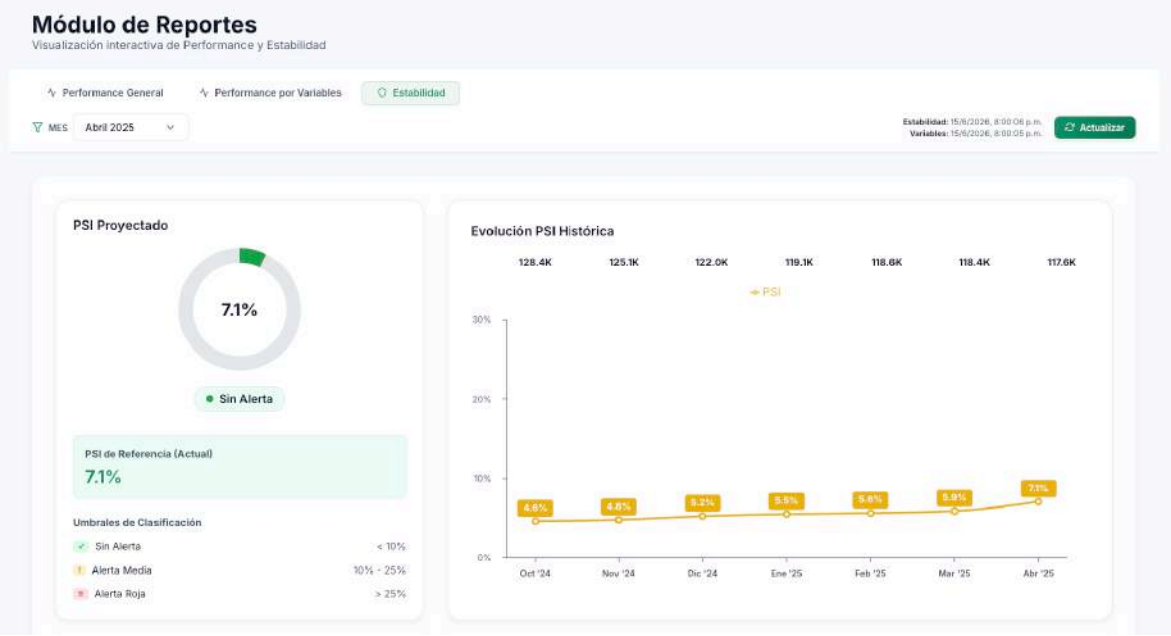


Nota. Fuente: Elaboración propia

En el frente de estabilidad, el score del modelo en su último monitoreo presenta un PSI del 7.1 %, lo que lo clasifica como un indicador sin alertas. Además, como se aprecia en la Figura 71, la evolución del PSI ha sido leve a lo largo de los meses. Esto indica que el modelo no ha sufrido cambios poblacionales fuertes respecto a su desarrollo.

Figura 71.

Vista Estabilidad General del Modelo disponible en la plataforma Model Risk Management Suite

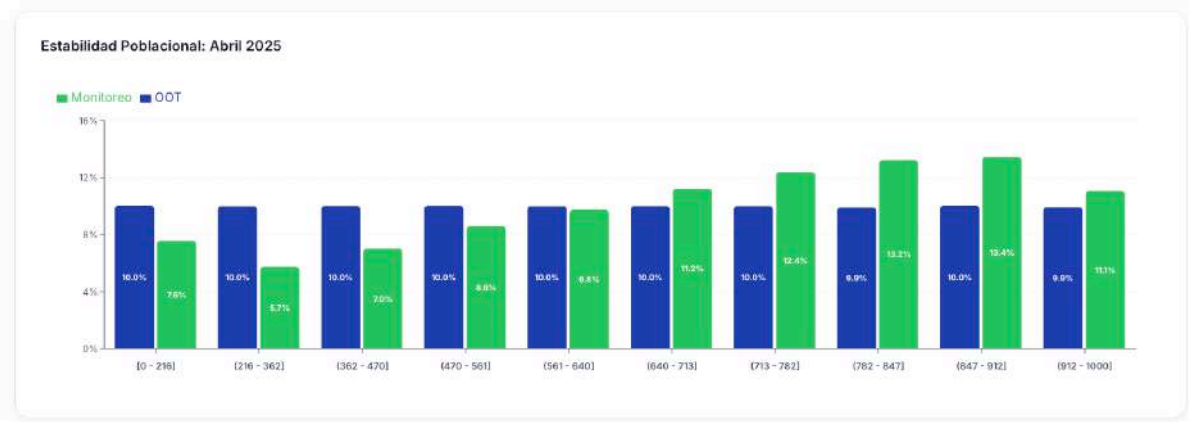


Nota. Fuente: Elaboración propia

Al comparar la distribución del score del último monitoreo con el OOT, se observa que la estabilidad poblacional ha mejorado. Los scores de mayor nivel (menos riesgosos) han crecido en comparación con el desarrollo, gracias a la gestión y uso del modelo, como lo muestra la Figura 72.

Figura 72.

Vista Estabilidad General del Modelo disponible en la plataforma Model Risk Management Suite



Nota. Fuente: Elaboración propia

Para finalizar la validación en estabilidad, se revisó el PSI por variables del top 11 de mayor importancia. En la Figura 73 se aprecia que las variables no han sufrido variaciones significativas de PSI. Solo una variable de tendencia ha mostrado ciertas variaciones, pero es un comportamiento normal. En la Figura 74 se observa que esta variable se ha desplazado más hacia la izquierda para sus valores no missing. Esto se traduce en que la población está decrementando su deuda, un buen indicador para la cartera de cualquier institución financiera.

Figura 73.

PSI por variables disponible en la plataforma Model Risk Management



Evolución de PSI por Variables

Peso	Variable	Nov '24	Dic '24	Ene '25	Feb '25	Mar '25	Abr '25
8.9%	% de Uso Total de Tarjeta de Crédito Último Mes	2.2%	2.0%	2.0%	1.8%	1.8%	1.8%
6.4%	Mínima Cantidad de Pasivos Total Últimos 3 Meses	3.6%	3.8%	4.0%	4.4%	4.3%	4.7%
4.5%	Máximo Días de Atraso Riesgo Directo Persona Natural Últimos 3 Meses	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
3.9%	Tendencia de Activos Totales Últimos 12 Meses	17.1%	16.8%	16.6%	16.0%	14.8%	13.6%
3.8%	Días de Atraso Riesgo Directo Persona Natural Último Mes	0.5%	0.3%	0.2%	0.3%	0.3%	0.3%
3.8%	Indicador de Cliente Dependiente	3.1%	3.7%	2.9%	4.2%	4.6%	4.7%
3.7%	Máximo Días de Atraso Riesgo Directo Persona Natural Último Año	1.6%	1.6%	1.3%	1.4%	1.3%	1.1%
3.2%	Promedio de Decrementos Último Año	3.5%	4.9%	5.3%	4.8%	5.8%	5.9%
3.0%	Ratio Pasivos Fin de Mes entre Pasivos Promedio Últimos 3 Meses	0.6%	0.8%	0.5%	0.5%	0.5%	1.4%
2.8%	Mínimo Pasivo Cuenta Sueldo Últimos 12 Meses	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%
2.7%	Número de Decrementos Último Año	7.7%	8.9%	9.3%	9.0%	9.8%	9.8%

Nota. Fuente: Elaboración propia

Figura 74.

Distribución de la Variable con Alerta Ambar disponible en la plataforma Model Risk Management



Nota. Fuente: Elaboración propia

Para evaluar la calidad de la plataforma desde la perspectiva de los usuarios finales, se aplicó un cuestionario de 5 preguntas basado en el modelo ISO/IEC 25010 a 3 integrantes del equipo de Model Risk Management, utilizando la escala Likert de 5 puntos que se observa en la Tabla 41.

Tabla 41.

Tabla Likert

Puntaje	Significado
1	Muy Deficiente
2	Deficiente
3	Aceptable
4	Bueno
5	Muy Bueno

Nota. Fuente: Elaboración propia

Los resultados para cada integrante (Para más detalle véase el Anexo F, G, H – Documento de Pruebas de Software) muestran que la plataforma obtuvo un puntaje dentro del rango BUENO – MUY BUENO en promedio para los elementos evaluados en la Tabla 42.

Tabla 42.

Criterios Evaluados

Aspecto	Pregunta
Usabilidad	La plataforma es fácil de usar.
Rendimiento	La plataforma responde rápidamente.
Funcionalidad	Las herramientas de consulta funcionan correctamente.
Seguridad	Confío en que la plataforma es segura.
Satisfacción	Estoy satisfecho con la plataforma.

Nota. Fuente: Elaboración propia

CAPÍTULO V: DISCUSIÓN, CONCLUSIONES Y RECOMENDACIONES

5.1 Discusión

El primer objetivo específico fue construir un dataset que integrara variables crediticias, sociales y comportamentales. Se partió de una base inicial de 2,235,642 clientes con préstamos personales vigentes entre junio de 2021 y octubre de 2022. Tras aplicar filtros de negocio y calidad, se obtuvo una población objetivo-limpia de 1,735,340 registros (1,530,737 buenos y 204,603 malos). La construcción del target mediante análisis de Roll Rates permitió identificar los 75 días de mora como punto de corte óptimo. A partir de ese umbral, la tasa de deterioro se estabiliza y la probabilidad de recuperación del cliente cae de forma sostenida.

Este enfoque coincide con lo que proponen Djeundje et al. (2021) y Hlongwane et al. (2024). Ambos autores señalan que la correcta definición del horizonte temporal y el punto de corte del target son decisiones fundacionales que condicionan todo el proceso de modelado posterior. La estabilidad de la tasa de malos a lo largo del periodo analizado que incluye tanto la crisis pandémica como el periodo de recuperación refuerza la representatividad de la muestra y la robustez esperada del modelo en escenarios adversos.

El segundo objetivo específico abordó el preprocesamiento del dataset. Las 582 variables iniciales, distribuidas en seis fuentes de información, fueron sometidas a un proceso escalonado de reducción de dimensionalidad. Este proceso combinó filtros por missings (> 95 %), Gini mínimo (< 5 %) y correlación (> 60 %), seguido de un análisis bivariado con transformación WOE y un toy model basado en importancia SHAP. Al final, el dataset quedó con 48 variables predictoras.

Esta estrategia de selección multietapa es consistente con los hallazgos de Muhammad Noor et al. (2024), quienes demuestran que la eliminación de variables altamente correlacionadas mejora el desempeño final del modelo. Asimismo, la transformación mediante Peso de la Evidencia (WOE) para variables categóricas como Procedencia Profesional y

Segmentación RFM siguió el enfoque recomendado por la literatura especializada en scoring crediticio. Permite capturar la relación monótona de cada variable con el target de manera interpretable. El tratamiento de outliers mediante Isolation Forest garantizó que la limpieza no introdujera sesgo por clase, al verificar que la distribución de buenos y malos fuera equivalente en los registros normales y anómalos.

El tercer objetivo específico fue seleccionar y entrenar el mejor modelo predictivo. El benchmarking entre siete algoritmos mostró que XGBoost obtuvo el mayor Gini en entrenamiento (76.3 %), aunque con una diferencia de 10.1 puntos frente al conjunto de test. Esto indicaba sobreajuste en su configuración inicial. Este comportamiento es conocido en la literatura. Yin et al. (2024) señalan que XGBoost puede sobreajustarse sin una configuración adecuada de sus hiperparámetros de regularización. Frente a ello, se optó por la optimización bayesiana de hiperparámetros mediante Optuna sobre Amazon SageMaker, utilizando el conjunto de datos completo de entrenamiento (más de 1.4 millones de registros).

El modelo final alcanzó un AUC de 84.15 % y un Gini de 68.29 % en el conjunto OOT, clasificado como "Bueno" según las escalas de Mittigová (2018). La diferencia promedio entre entrenamiento y validación fue de apenas 0.3 puntos de AUC, lo que confirma que el sobreajuste fue controlado. Estos resultados son comparables con los de Aruleba y Sun (2024), quienes reportan AUC de hasta 0.87 con XGBoost en conjuntos de datos crediticios usando SMOTE-ENN. También coinciden con los de González Manrique y Coveñas De León (2024), quienes obtuvieron un AUC de 0.95 con XGBoost en el contexto peruano de inclusión financiera, aunque sobre un dataset de características distintas. El estadístico KS de 0.51 en OOT, considerado de "Buena discriminación" según Barreto (2025), complementa la evidencia de que el modelo separa adecuadamente a los clientes buenos de los malos en una ventana de 12 meses.

Vale la pena analizar por qué XGBoost fue seleccionado sobre LightGBM, que obtuvo un Gini OOT ligeramente superior en el benchmarking inicial (68.9 % vs 66.2 %). Vásquez (2025) y Aradiel Castañeda et al. (2025) reportan que LightGBM supera a XGBoost en precisión pura en ciertos contextos peruanos. Sin embargo, la brecha entre ambos modelos se redujo a menos de 1 punto de Gini tras la optimización bayesiana de XGBoost. Su capacidad de regularización a través de los parámetros gamma, lambda y alpha, y el manejo explícito del desbalance de clases mediante scale_pos_weight ofrecen ventajas adicionales en términos de estabilidad del modelo en producción. Este resultado concuerda con lo que plantean Alonso Robisco y Carbó Martínez (2022): en entornos bancarios regulados, la estabilidad y la explicabilidad son componentes tan relevantes como la precisión puntual.

El cuarto objetivo específico fue implementar un sistema de seguimiento de performance, estabilidad e interpretabilidad. El análisis mediante valores SHAP reveló que el porcentaje de uso de tarjeta de crédito en el último mes (pct_uso_tctot_u1m) es la variable de mayor peso relativo (8.9 %). Le siguen el mínimo de pasivos totales en los últimos 3 meses (6.4 %) y el máximo de días de atraso en el sistema financiero en los últimos 3 meses (4.5 %). Este hallazgo tiene una interpretación económica clara: el comportamiento de consumo sobre la línea disponible y la capacidad de mantener saldo pasivo son señales tempranas de estrés financiero que preceden al default. Liu et al. (2024) obtienen resultados similares al aplicar SHAP sobre modelos de ensamble, donde las variables de uso crediticio y comportamiento de pago reciente tienden a concentrar el peso explicativo.

Por otra parte, el Force Plot confirma que el modelo opera de forma coherente a nivel individual. Los empujes positivos y negativos sobre el log-odds base son interpretables y alineados con el sentido económico de cada variable. Esto facilita su adopción en procesos de decisión crediticia y su explicabilidad ante la SBS, conforme a los requerimientos de la Resolución N° 00053-2023.

El monitoreo en producción de la plataforma Model Risk Management Suite mostró que el modelo mantuvo un Gini de casi 70 puntos en su último mes de seguimiento (abril de 2025), 1.4 puntos por encima del OOT original. El PSI del score fue de 7.1 %, sin alertas, lo que indica estabilidad poblacional. Solo una variable de tendencia mostró un desplazamiento hacia la izquierda en su distribución, lo que refleja que la cartera ha reducido su endeudamiento. Ese es un comportamiento positivo desde la perspectiva del riesgo crediticio.

Estos resultados son consistentes con las recomendaciones de De Lange et al. (2022), quienes destacan que los modelos de gradient boosting explicados con SHAP pueden sostenerse en producción sin perder la transparencia exigida por los supervisores. La validación independiente realizada por una consultora externa, que evaluó replicabilidad del código, solidez metodológica y bondad de ajuste, ratificó que el modelo cumple los criterios mínimos para su puesta en producción. Esto concuerda con el marco de Gestión de Riesgos de Modelo de la SBS.

Entre las limitaciones de la investigación destacan tres. Primera, la población objetivo está acotada a clientes con préstamos personales vigentes en una única entidad bancaria. Por lo tanto, los resultados no son directamente extrapolables a otros productos crediticios ni a otras instituciones financieras. Segunda, la tasa de malos resultante (11.8 %) puede diferir de la que enfrente el modelo en periodos económicos distintos. Por ello, el scale_pos_weight deberá recalibrarse si cambia significativamente la distribución de impago. Tercera, el análisis de fairness, decir la evaluación de si el modelo discrimina de forma diferencial entre grupos sociodemográficos no fue abordado en este estudio. Constituye un área pendiente de importancia creciente según Hurlin et al. (2026) y Moldovan (2023), especialmente ante las exigencias regulatorias emergentes sobre equidad algorítmica en el sector financiero.

5.2. Conclusiones

5.2.1. Conclusión general

Objetivo general: Desarrollar un Modelo de Machine Learning usando el algoritmo XGBOOST para medir la probabilidad de impago de clientes en una entidad bancaria, 2026

Conclusión general: La investigación desarrolló una solución para evaluar el riesgo de impago de clientes con préstamos personales en la entidad bancaria estudiada. La solución abarca la construcción y preparación de los datos, un modelo XGBoost, herramientas de interpretación basadas en SHAP y una plataforma para consultar indicadores de desempeño y estabilidad. En la evaluación fuera del tiempo, correspondiente a septiembre y octubre de 2022, el modelo obtuvo un AUC de 84.15 %, un Gini de 68.29 % y un KS de 0.51. Estos resultados respaldan su capacidad para discriminar entre los grupos definidos en la investigación durante dicho periodo. El estudio también documenta la implementación del monitoreo y la revisión independiente del artefacto. Los resultados corresponden a la cartera de préstamos personales, la entidad y los periodos analizados; su aplicación a otras poblaciones requiere una evaluación adicional.

5.2.2. Conclusiones específicas

Objetivo específico 1: Identificar las variables de carácter crediticio, social y comportabilidad para la construcción del dataset inicial que sirva para medir la probabilidad de impago de clientes en una entidad bancaria, 2026.

Conclusión específica 1: Se identificaron e integraron 582 variables provenientes de seis fuentes de información: uso crediticio (305), parque financiero (96), productos (84), información sociodemográfica (42), información móvil (27) y propiedad (28), según la Tabla 11. Estas fuentes aportaron datos sobre utilización del crédito, obligaciones en el sistema financiero, tenencia de productos y características de los clientes. Con ello se construyó el conjunto de datos inicial para estudiar el comportamiento de pago de la cartera de préstamos

personales. Registrar cuántas variables aportó cada fuente permitió documentar la procedencia de los predictores antes de su selección y transformación.

Objetivo específico 2: Preprocesar el dataset inicial que va a servir para medir la probabilidad de impago de clientes en una entidad bancaria, 2026.

Conclusión específica 2: Se aplicó un proceso escalonado de preparación y selección de variables. Primero se retiraron las que superaban el 95 % de valores faltantes, las que presentaban un Gini inferior al 5 % y las que excedían el criterio de correlación del 60 %, conforme a los filtros descritos en el apartado 4.1.2. Luego se realizó el análisis bivariado, se trataron los valores faltantes y se transformaron variables categóricas mediante WOE. La selección se completó con un modelo de prueba y un análisis de importancia, tras los cuales quedaron 48 variables predictoras, como se indica en la p. 88. El procedimiento también incluyó la detección de valores atípicos mediante Isolation Forest y la construcción de la variable objetivo a partir del análisis de roll rates para el horizonte de evaluación establecido. Así se obtuvo el conjunto de datos utilizado en el entrenamiento del modelo final.

Objetivo específico 3: Entrenar el mejor Modelo de Machine Learning para medir probabilidad de impago de clientes en una entidad bancaria, 2026.

Conclusión específica 3: Se evaluaron siete algoritmos y, tras la etapa de comparación descrita en la investigación, se entrenó y optimizó un modelo XGBoost mediante Optuna en Amazon SageMaker. Para su desarrollo se utilizaron los periodos de junio de 2021 a agosto de 2022, con validación cruzada estratificada de cinco pliegues; septiembre y octubre de 2022 se reservaron como periodo fuera del tiempo, según la Tabla 33. El AUC promedio fue de 84.57 % en entrenamiento y 84.29 % en validación, una diferencia de 0.28 puntos porcentuales. En el conjunto OOT, el modelo alcanzó un AUC de 84.15 %, un Gini de 68.29 % y un KS de 0.51 (Tabla 35 y apartado 4.1.3.5). Estas métricas reflejan la capacidad discriminante del modelo en la muestra y en el periodo de evaluación indicados.

Objetivo específico 4: Implementar el monitoreo del Modelo que mide la probabilidad de impago de clientes en una entidad bancaria, 2026.

Conclusión específica 4: Se implementó un flujo de cálculo y consulta de indicadores de desempeño y estabilidad en la plataforma Model Risk Management Suite. La arquitectura descrita en el apartado 4.1.4 utiliza AWS Step Functions para coordinar los procesos, AWS Glue para calcular las métricas y S3 y Athena para almacenar y consultar los resultados. La interpretación del modelo se incorporó mediante SHAP: las tres variables con mayor peso relativo registraron 8.9 %, 6.4 % y 4.5 %, respectivamente, y sumaron 19.8 % del peso presentado en la Tabla 39. En los tableros de abril de 2025 se muestran un Gini proyectado de 69.7 % y un PSI proyectado de 7.1 %, este último sin alerta según la clasificación de la plataforma (Figuras 67 y 71). La tesis documenta, además, una revisión externa del artefacto del modelo y una evaluación de la plataforma por integrantes del equipo de seguimiento y validación.

5.3. Recomendaciones

5.3.1. Recomendaciones para la entidad bancaria

Se recomienda implementar el score crediticio generado por el modelo como insumo transversal en las políticas de gestión de cartera activa. El valor predictivo del modelo en una ventana de 12 meses lo hace especialmente útil para decisiones de renovación de productos, reprogramación preventiva de deudas y segmentación de campañas de cobranza temprana. Esas son las áreas donde la anticipación del impago tiene mayor impacto económico.

También se sugiere establecer un proceso formal de recalibración del parámetro scale_pos_weight ante cambios estructurales en la tasa de malos. Ciclos económicos adversos o cambios normativos pueden alterar la proporción de clientes buenos y malos en la

cartera. Disponer de un protocolo de recalibración documentado reduce el riesgo de que el modelo pierda poder discriminante ante esas variaciones.

Otro punto importante es realizar una calibración del modelo analizando la tasa de probabilidad de default vs la tasa de malos real, de esta manera se podrá ver si el score del modelo se encuentra calibrado. En caso el modelo no se encuentre calibrado, se deben implementar metodologías para tener un score correcto y no afectar a cálculos de provisiones y magnitud.

Se propone integrar el Agente MRM desarrollado en esta investigación como herramienta de primera línea para los equipos de seguimiento de modelos. La capacidad del agente para responder consultas sobre indicadores de performance, estabilidad e interpretabilidad en lenguaje natural reduce los tiempos de análisis y democratiza el acceso a la información del modelo entre perfiles técnicos y de negocio.

Finalmente, se recomienda extender el enfoque metodológico empleado en esta investigación a otros productos de la cartera minorista como tarjetas de crédito, créditos vehiculares, microcréditos adaptando la definición del target y el horizonte temporal según las características de mora de cada producto. La infraestructura en la nube y el pipeline de CRISP-ML(Q) implementados son reutilizables con modificaciones menores.

5.3.2 Recomendaciones para futuras investigaciones

Para futuras investigaciones, se sugiere incorporar variables de datos alternativos información de telecomunicaciones, datos de movilidad georreferenciada o variables psicométricas para mejorar la capacidad predictiva en segmentos de la cartera con historial crediticio limitado. Djeundje et al. (2021) y Hlongwane et al. (2024) demuestran que estas variables pueden elevar el AUC hasta 0.79 en poblaciones con baja bancarización, un contexto relevante en el mercado peruano.

También se recomienda evaluar el desempeño diferencial del modelo por grupos sociodemográficos, género, nivel educativo, edad, región geográfica aplicando marcos formales de equidad algorítmica como los propuestos por Hurlin et al. (2026) y Moldovan (2023). Esta evaluación es necesaria para cumplir con estándares de equidad regulatoria emergentes y para garantizar que el modelo no introduzca ni amplifique sesgos en el acceso al crédito.

Se sugiere comparar el desempeño del modelo XGBoost optimizado frente a arquitecturas de aprendizaje profundo redes LSTM o Transformers adaptados a series temporales financieras en el mismo dataset, con el fin de cuantificar la ganancia predictiva que ofrecen los modelos más complejos versus el costo adicional en interpretabilidad y gobernanza de modelos. Perales (2024) reporta precisiones de 91.5 % con redes neuronales multicapa en cajas financieras peruanas, lo que sugiere que existe margen de mejora en métricas puntuales.

Se recomienda explorar técnicas de sobremuestreo más avanzadas como NOTE (Non-Parametric Oversampling Technique), propuesta por Han et al. (2024) como alternativa al SMOTE tradicional. Este enfoque puede reducir el solapamiento entre clases en datasets con alta proporción de buenos, mejorando la estabilidad de las predicciones en los percentiles superiores del score donde el modelo actualmente muestra mayor varianza.

Por último, se sugiere analizar la aplicabilidad del marco regulatorio SBS (Resolución N° 00053-2023) como estándar de gobernanza de modelos en instituciones financieras de menor tamaño cajas municipales, cooperativas, fintechs identificando las brechas de capacidad técnica y los costos de implementación de sistemas de monitoreo continuo como el desarrollado en esta investigación. Este análisis aportaría evidencia práctica sobre la escalabilidad del enfoque propuesto en el ecosistema financiero peruano.

REFERENCIAS

Abi, R. (2025). Machine learning for credit scoring and loan default prediction using behavioral and transactional financial data. *World Journal of Advanced Research and Reviews*, *26*(3), 884-904. https://doi.org/10.30574/wjarr.2025.26.3.2266

Agarwal, A. (2025). DC-Guard: A Risk-Tiered Data-Centric Framework for Governing Dataset Shift in Credit-Risk Models — Empirical Validation on Fannie Mae Loans. *Innovative Journal of Applied Science*, *02*(06). https://doi.org/10.70844/ijas.2025.2.45

Albónico, M. P. (2021). Herramientas de Machine Learning Aplicadas al Cálculo de Efectos de Tratamiento en Campañas de Marketing.

Alizade, İ. M. (2025). EXPLORING CREDIT RISKS AND RISK MANAGEMENT IN BUSINESS: A FINANCIAL PERSPECTIVE. *Tikintinin iqtisadiyyatı və menecment*, *2*, 12-18. https://doi.org/10.58225/tim.2025-2-12-18

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, *97*, 00133. https://doi.org/10.1051/bioconf/20249700133

Alonso Robisco, A., & Carbó Martínez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, *8*(1), 70. https://doi.org/10.1186/s40854-022-00366-1

Aradiel Castañeda, H., Mas Azahuanche, G. A., Mendoza Arenas, R. D., Castillo Paredes, O. T. A., Reinoso Palacios, A. R., Delgado Baltazar, M. P., & Mendoza Delgado, R. S. (2025). Explainable Machine Learning for Credit Card Default Prediction Using Web-Scraped Financial Data: A Case Study in the Peruvian Banking Sector. *Proceedings of the 5th LACCEI International Multiconference on Entrepreneurship, Innovation and Regional Development (LEIRD 2025): «Entrepreneurship with Purpose: Social and Technological Innovation in the Age of AI»*. 5th LACCEI

International Multiconference on Entrepreneurship, Innovation and Regional Development (LEIRD 2025): «Entrepreneurship with Purpose: Social and Technological Innovation in the Age of AI». https://doi.org/10.18687/LEIRD2025.1.1.451

Arini, F. Y., Djuanda, L. A., Kristianto, A. H. P., Tiadah, M. N., Wicaksono, A. P., & Putra, F. A. A. (2026). Performance Evaluation of Gradient Boosting Techniques for Predicting Customer Purchase Decisions. *Jurnal Teknik Informatika (Jutif)*, *7*(2), 1441-1454. https://doi.org/10.52436/1.jutif.2026.7.2.5461

Aruleba, I., & Sun, Y. (2024). Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation. *IEEE Access*, *12*, 115015-115025. https://doi.org/10.1109/ACCESS.2024.3445308

Balboa, A., Cuesta, A., González-Villa, J., Ortiz, G., & Alvear, D. (2024). Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations. *Safety Science*, *174*, 106485. https://doi.org/10.1016/j.ssci.2024.106485

Barreto, M. F. (2025). *Publicación: Diseño de un modelo predictivo para la optimización temporal de cobranza mediante técnicas de machine learning* [Tesis de pregrado, Escuela Colombiana de Ingeniería Julio Garavito.]. https://repositorio.escuelaing.edu.co/handle/001/3789

Bhandari, B. P., Dhakal, S., & Tsou, C.-Y. (2024). Assessing the Prediction Accuracy of Frequency Ratio, Weight of Evidence, Shannon Entropy, and Information Value Methods for Landslide Susceptibility in the Siwalik Hills of Nepal. *Sustainability*, *16*(5), 2092. https://doi.org/10.3390/su16052092

Boggavarapu, V. (2025). *Innovations in DataOps for AI in Financial Services: Automation and Efficiency*. https://doi.org/10.5281/ZENODO.16872258

Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, *15*(1), 73. https://doi.org/10.1186/s13321-023-00743-7

Bounab, R., Zarour, K., Guelib, B., & Khlifa, N. (2024). Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN. *IEEE Access*, *12*, 54382-54396. https://doi.org/10.1109/ACCESS.2024.3385781

Brygała, M., & Korol, T. (2024). Personal bankruptcy prediction using machine learning techniques. *Economics and Business Review*, *10*(2), 118-142. https://doi.org/10.18559/ebr.2024.2.1149

Calin, A. M., Pargaru, O. I., Raducan, D., & Buzoianu, O. A. C. (2023). Determinants of the Probability of Non-repayment of Loans Granted to Individuals. *European Journal of Sustainable Development*, *12*(4), 217. https://doi.org/10.14207/ejsd.2023.v12n4p217

Campanioni, S., Veiga, C., Prieto-González, J. M., González-Nóvoa, J. A., Busto, L., Martinez, C., Alberte-Woodward, M., García De Soto, J., Pouso-Diz, J., Fernández Ceballos, M. D. L. Á., & Agis-Balboa, R. C. (2024). Explainable machine learning on baseline MRI predicts multiple sclerosis trajectory descriptors. *PLOS ONE*, *19*(7), e0306999. https://doi.org/10.1371/journal.pone.0306999

Canbek, G., Taskaya Temizel, T., & Sagiroglu, S. (2022). PToPI: A Comprehensive Review, Analysis, and Knowledge Representation of Binary Classification Performance Measures/Metrics. *SN Computer Science*, *4*(1), 13. https://doi.org/10.1007/s42979-022-01409-1

Chan, C.-P., Tang, F.-K., Tsai, C.-H., & Yang, J.-H. (2025). A SHAP-Based Comparative Analysis of Machine Learning Model Interpretability in Financial Classification

Tasks. *Journal of Applied Economic Sciences (JAES)*, *20*(16), 385. https://doi.org/10.57017/jaes.v20.3(89).03

Chen, R., Wang, S., Zhu, Z., Yu, J., & Dang, C. (2023). Credit ratings of Chinese online loan platforms based on factor scores and K-means clustering algorithm. *Journal of Management Science and Engineering*, *8*(3), 287-304. https://doi.org/10.1016/j.jmse.2022.12.003

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357-372. https://doi.org/10.1016/j.ejor.2023.06.036

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

Das, P., Ivkin, N., Bansal, T., Rouesnel, L., Gautier, P., Karnin, Z., Dirac, L., Ramakrishnan, L., Perunicic, A., Shcherbatyi, I., Wu, W., Zolic, A., Shen, H., Ahmed, A., Winkelmolen, F., Miladinovic, M., Archembeau, C., Tang, A., Dutt, B., … Venkateswar, K. (2020). Amazon SageMaker Autopilot: A white box AutoML solution at scale. *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, 1-7. https://doi.org/10.1145/3399579.3399870

Dasril, Y., Muslim, M. A., Hakim, M. F. A., Jumanto, J., & Prasetiyo, B. (2023). Credit Risk Assessment in P2P Lending Using LightGBM and Particle Swarm Optimization. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *9*(1), 18-28. https://doi.org/10.26594/register.v9i1.3060

De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Dembélé, D. (2025). *The Eigenvalues Entropy as a Classifier Evaluation Measure* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2511.01904

Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, *163*, 113766. https://doi.org/10.1016/j.eswa.2020.113766

Egwa, A. A., Kakudi, H. A., Ahmad, A. A., Bichi, A. M., & Madu, M. A. (2022). Prediction Model for Loan Default Using Machine Learning. *The International Journal of Science & Technoledge*, *10*(2). https://doi.org/10.24940/theijst/2022/v10/i2/ST2202-009

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903-4923. https://doi.org/10.1007/s10994-022-06296-4

Foglia, M. (2022). Non-Performing Loans and Macroeconomics Factors: The Italian Case. *Risks*, *10*(1), 21. https://doi.org/10.3390/risks10010021

Giraldo, W. O., & Marín, J. (2021). Machine Learning para la estimación del riesgo de crédito en una cartera de consumo. Universidad EAFIT.

Gonzalez Manrique, P. M., & Coveñas De León, M. (2024). Construcción de un Score Crediticio Mediante el Análisis de Datos Alternativos y Métodos de Inteligencia Artificial para impulsar la Inclusión Financiera. *Revista de Economía y Finanzas*, *2*(5). https://doi.org/10.32826/reyf.v2i5.373

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3.ª ed.). Morgan Kaufmann.

Han, S., & Jung, H. (2024). NATE: Non-pArameTric approach for Explainable credit scoring on imbalanced class. *PLOS ONE*, *19*(12), e0316454. https://doi.org/10.1371/journal.pone.0316454

Han, S., Jung, H., Yoo, P. D., Provetti, A., & Cali, A. (2024). NOTE: Non-parametric oversampling technique for explainable credit scoring. *Scientific Reports*, *14*(1), 26070. https://doi.org/10.1038/s41598-024-78055-5

Hermitaño Castro, J. A. (2022). Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática. *Interfases*, (015), 160-178. https://doi.org/10.26439/interfases2022.n015.5898

Hernández Sampieri, R., & Fernandez-Collado, C. F. (2014). *Metodología de la investigación* (P. Baptista Lucio, Ed.; Sexta edición). McGraw-Hill Education.

Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE*, *19*(5), e0303566. https://doi.org/10.1371/journal.pone.0303566

Hsu, T.-C., Tsai, Y.-H., Chu, W. C.-C., Chen, S., Tsai, H.-L., & Chang, Y.-K. (2023). Exploration of advanced computer technology to address analytical and noise improvement issues in machine learning. *Journal of Systems and Software*, *205*, 111820. https://doi.org/10.1016/j.jss.2023.111820

Hurlin, C., Pérignon, C., & Saurin, S. (2026). The Fairness of Credit Scoring Models. *Management Science*, *72*(1), 406-425. https://doi.org/10.1287/mnsc.2022.03888

Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, *18*(1), 37. https://doi.org/10.3390/a18010037

Infobae. (2022, diciembre 27). Resumen de productos financieros más buscados por los peruanos. *Infobae*. https://www.infobae.com/america/peru/2022/12/27/resumen-de-productos-financieros-mas-buscados-por-los-peruanos/

Japinye, A. O., & Adedugbe, A. A. (2025). *Explainable AI for Credit Scoring with SHAP-Calibrated Ensembles: A Multi-Market Evaluation on Public Lending Data*. https://doi.org/10.5281/ZENODO.17155174

John, S. (2025). *Fair and Explainable Credit-Scoring under Concept Drift: Adaptive Explanation Frameworks for Evolving Populations* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2511.03807

Joshi, S. (2025). *AI and Financial Model Risk Management: Applications, Challenges, Explainability, and Future Directions*. Computer Science and Mathematics. https://doi.org/10.20944/preprints202503.2284.v1

Khademi, A., Hopka, M., & Upadhyay, D. (2023). *Model Monitoring and Robustness of In-Use Machine Learning Models: Quantifying Data Distribution Shifts Using Population Stability Index* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2302.00775

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309-323. https://doi.org/10.1016/j.ejor.2021.12.024

Kurian, J. F., & Allali, M. (2024). Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications. *Journal of Data, Information and Management*, *6*(3), 207-216. https://doi.org/10.1007/s42488-024-00119-y

Kyeong, S., & Shin, J. (2022). Two-stage credit scoring using Bayesian approach. *Journal of Big Data*, *9*(1), 106. https://doi.org/10.1186/s40537-022-00665-5

Lavazza, L., & Morasca, S. (2023). Common Problems With the Usage of F-Measure and Accuracy Metrics in Medical Research. *IEEE Access*, *11*, 51515-51526. https://doi.org/10.1109/ACCESS.2023.3278996

Li, Y., & Sha, M. (2024). Two-stage credit risk prediction framework based on three-way decisions with automatic threshold learning. *Journal of Forecasting*, *43*(5), 1263-1277. https://doi.org/10.1002/for.3074

Liang, J., Sugang, Lin, H., Wu, Y., Zhao, R., & Li, Z. (2025). Reasoning RAG via System 1 or System 2: A Survey on Reasoning Agentic Retrieval-Augmented Generation for Industry Challenges. *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 1954-1966. https://doi.org/10.18653/v1/2025.findings-ijcnlp.122

Lin, L., & Wang, Y. (2025). SHAP Stability in Credit Risk Management: A Case Study in Credit Card Default Model. *Risks*, *13*(12), 238. https://doi.org/10.3390/risks13120238

Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (2024). Credit scoring prediction leveraging interpretable ensemble learning. *Journal of Forecasting*, *43*(2), 286-308. https://doi.org/10.1002/for.3033

López, J. C. (2019). *Comparación de modelos de aprendizaje de máquina en la predicción del incumplimiento de pago en el sector de las microfinanzas* [PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ]. https://tesis.pucp.edu.pe/items/ff4eaa73-f4f7-4fe0-9b29-7ea413ab2a56

Lovell, D., Miller, D., Capra, J., & Bradley, A. (2022). *Never mind the metrics -- what about the uncertainty? Visualising confusion matrix metric distributions* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2206.02157

Magdaong, J. B., Culaba, A. B., Ubando, A. T., & Lopez, N. S. (2024). Generating synthetic building electrical load profiles using machine learning based on the CRISP-ML(Q) framework. *IOP Conference Series: Earth and Environmental Science*, *1372*(1), 012082. https://doi.org/10.1088/1755-1315/1372/1/012082

Martinez Preciado, C. J., Arbulu Anicama, J. M., Arotinco Gomez, W. M., Belloda Saavedra, C. E., & Jorge Salazar, A. (2024). *Modelo de Segmentación de Riesgo de Lavado de Activos aplicado a personas naturales en una microfinanciera según regulación de la SBS, 2024* [Universidad Peruana de Ciencias Aplicadas]. http://hdl.handle.net/10757/685991

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Mestiri, S. (2024). Credit scoring using machine learning and deep Learning-Based models. *Data Science in Finance and Economics*, *4*(2), 236-248. https://doi.org/10.3934/DSFE.2024009

Metawa, N., Afchal, S., & El-Kanj, N. (2025). AI-based model for Enhancing Credit Risk and Delinquency Management in Banks. *Journal of Cybersecurity and Information Management*, *16*(1). https://doi.org/10.54216/JCIM.160120

Mienye, I. D., & Sun, Y. (2022a). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

Mienye, I. D., & Sun, Y. (2022b). Effective Feature Selection for Improved Prediction of Heart Disease. En T. M. N. Ngatched & I. Woungang (Eds.), *Pan-African Artificial Intelligence and Smart Systems* (Vol. 405, pp. 94-107). Springer International Publishing. https://doi.org/10.1007/978-3-030-93314-2_6

Mittigová, P. (2018). *Consumer Credit Risk Analysis: Evidence from the Czech Republic* [Univerzita Karlova, Fakulta sociálních věd]. http://hdl.handle.net/20.500.11956/99623

Miyaji, R. O., & Corrêa, P. L. P. (2025). Enhancing ML Explainability with Multi-Agent LLMs: A Context-Aware XAI Approach. *Anais do XL Simpósio Brasileiro de Banco de Dados (SBBD 2025)*, 28-41. https://doi.org/10.5753/sbbd.2025.246998

Moldovan, D. (2023). Algorithmic Decision Making Methods for Fair Credit Scoring. *IEEE Access*, *11*, 59729-59743. https://doi.org/10.1109/ACCESS.2023.3286018

Muhammad Noor, Radityo Adi Nugroho, Setyo Wahyu Saputro, Rudy Herteno, & Friska Abadi. (2024). Optimization of Backward Elimination for Software Defect Prediction with Correlation Coefficient Filter Method. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(4), 397-404. https://doi.org/10.35882/jeeemi.v6i4.466

Muñoz, J., Efthimiou, O., Audigier, V., De Jong, V. M. T., & Debray, T. P. A. (2024). Multiple imputation of incomplete multilevel data using Heckman selection models. *Statistics in Medicine*, *43*(3), 514-533. https://doi.org/10.1002/sim.9965

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023). On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance. *Expert Systems with Applications*, *218*, 119599. https://doi.org/10.1016/j.eswa.2023.119599

Naidu, Y. P. K., & Amanullah, M. (2025). *An efficient air pollution prediction system using enhanced CatBoost classifier compared over AdaBoost classifier with improved accuracy*. 020147. https://doi.org/10.1063/5.0258921

Nazir, A. A., Saiful, I., & Sohan, Md. F. A. A. (2025). Machine Learning Models for Credit Risk Assessment. *2025 IEEE International Conference on Signal Processing,*

Information, Communication and Systems (SPICSCON), 307-312. https://doi.org/10.1109/SPICSCON69221.2025.11504159

Nguyen, Q. G., Nguyen, L. H., Hosen, M. M., Rasel, M., Shorna, J. F., Mia, M. S., & Khan, S. I. (2025). Enhancing credit risk management with machine learning: A comparative study of predictive models for credit default prediction. *The American Journal of Applied Sciences*, *7*(1), 21–30. https://doi.org/10.37547/tajas/Volume07Issue01-04

Nithya, K., Mahesh, P., Reddy, B. R., & Ramakrishna, P. (2025). Forecasting Rainfall Patterns With Random Forest Algorithm. *2025 IEEE 6th India Council International Subsections Conference (INDISCON)*, 1-6. https://doi.org/10.1109/INDISCON66021.2025.11252303

Ñaupas Paitán, H., Mejía Mejía, E., Novoa Ramírez, E., & Villagómez Paucar, A. (2014). *Metodología de la investigación: Cuantitativa, cualitativa y redacción de la tesis* (4 ed). Ediciones de la U.

Paz, H. R. (2024). Factores del Abandono Universitario: Análisis con LightGBM y la Teoría de Juegos Cooperativo de Shapley. Revista Educación Superior Y Sociedad (ESS), 36(1), 163-193. https://doi.org/10.54674/ess.v36i1.872

Paladugu, N. (2023). Large Language Models for Automated Financial Code Generation and Documentation in Data Pipelines. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(1). https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P113

Perales, H. A. (2024). *Modelo de calificación basado en redes neuronales para la predicción de riesgo crediticio: Caja financiera* [Tesis de pregrado, Universidad de Lima]. https://alicia.concytec.gob.pe/vufind/Record/RULI_29b5df3db850d159e4ad3346e12dc05c

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11), e70056. https://doi.org/10.1111/cts.70056

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics andstatistical tests for machine learning. *Scientific Reports*, *14*(1), 15724. https://doi.org/10.1038/s41598-024-66611-y

Ramakrishna, M. T., Venkatesan, V. K., Izonin, I., Havryliuk, M., & Bhat, C. R. (2023). Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data. *Entropy*, *25*(2), 245. https://doi.org/10.3390/e25020245

Rao, C., Liu, Y., & Goh, M. (2023). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. *Complex & Intelligent Systems*, *9*(2), 1391-1414. https://doi.org/10.1007/s40747-022-00854-y

Gestión. (2021). Morosidad en tarjetas de créditos supera al doble de los préstamos personales. *Gestión*. https://gestion.pe/tu-dinero/finanzas-personales/morosidad-en-tarjetas-de-creditos-supera-al-doble-de-los-prestamos-personales-noticia/

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, *5*(6), 100994. https://doi.org/10.1016/j.patter.2024.100994

Saliba, C., Farmanesh, P., & Athari, S. A. (2023). Does country risk impact the banking sectors' non-performing loans? Evidence from BRICS emerging economies. *Financial Innovation*, *9*(1), 86. https://doi.org/10.1186/s40854-023-00494-2

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. https://doi.org/10.1002/aisy.202400304

Sarkar, R. (2026). A Systematic Review of AI-Driven Credit Risk Assessment Models in Commercial Banking (2018–2026). *American Journal of Interdisciplinary Studies*, *07*(01), 459-495. https://doi.org/10.63125/m52yna23

Scott, A. O., Prisca Amajuoyi, & Kudirat Bukola Adeusi. (2024). Advanced risk management solutions for mitigating credit risk in financial operations. *Magna Scientia Advanced Research and Reviews*, *12*(1), 212-223. https://doi.org/10.30574/msarr.2024.11.1.0085

Seitshiro, M. B., & Govender, S. (2024). Credit risk prediction with and without weights of evidence using quantitative learning models. *Cogent Economics & Finance*, *12*(1), 2338971. https://doi.org/10.1080/23322039.2024.2338971

Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised Machine Learning: Algorithms and Applications. En P. Singh (Ed.), *Fundamentals and Methods of Machine and Deep Learning* (1.ª ed., pp. 1-16). Wiley. https://doi.org/10.1002/9781119821908.ch1

Shinagam, R. K., Maruvada, T., Janjeti, S., Talari, R. M. D., & Shinagam, R. V. (2024). Development of a Machine Learning Algorithm for Efficient Localization of Damage in a Composite Structure Using Random Forest Technique. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, *48*(6), 4793-4809. https://doi.org/10.1007/s40996-024-01469-8

Shu, Y., Liu, Y., & Xie, Z. (2026). *Agentic Retrieval-Augmented Generation for Financial Document Question Answering* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2605.05409

Si, Z., Niu, H., & Wang, W. (2022). Credit Risk Assessment by a Comparison Application of Two Boosting Algorithms. En A. J. Tallón-Ballesteros (Ed.), *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA220367

Slabber, E., Verster, T., & De Jongh, R. (2023). Some Insights about the Applicability of Logistic Factorisation Machines in Banking. *Risks*, *11*(3), 48. https://doi.org/10.3390/risks11030048

Stacyana, J. S., & Chairunisah. (2026). *Comparison of Logistic Regression and XGBoost Model Performance in Predicting Credit Scores*. https://mail.ioinformatic.org/index.php/JAIEA/article/view/1877

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392-413. https://doi.org/10.3390/make3020020

Sudjianto, A., & Liu, A. J. (2025). *Decomposing Global AUC into Cluster-Level Contributions for Localized Model Diagnostics* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2508.07495

Superintendencia de Banca, Seguros y AFP. (2023). *Reglamento de Gestión de Riesgos de Modelo*. https://intranet2.sbs.gob.pe/dv_int_cn/2240/v1.0/Adjuntos/0053-2023.R.pdf

Suresh Dameruppula. (2025). The evolution and impact of high-performance computing systems. *World Journal of Advanced Research and Reviews*, *26*(1), 1983-1989. https://doi.org/10.30574/wjarr.2025.26.1.1247

Suxo, G. W. (2024). *Implementación de una solución de machine learning para la asignación de créditos financieros en una entidad financiera* [Trabajo de Suficiencia Profesional, Universidad Nacional Mayor de San Marcos]. https://alicia.concytec.gob.pe/vufind/Record/UNMS_20b0b4a65a65bf120eb323face50381e

Tigges, M., Mestwerdt, S., Tschirner, S., & Mauer, R. (2024). Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data. *Technological Forecasting and Social Change*, *205*, 123491. https://doi.org/10.1016/j.techfore.2024.123491

Toledo Concha, E. (2022). Factores que influyen en el riesgo de crédito de instituciones microfinancieras peruanas: Cajas Municipales de Ahorro y Crédito. *360: Revista de Ciencias de la Gestión*, (7). https://doi.org/10.18800/360gestion.202207.008

Tyler, M., & McCeney, J. (2024). Assured AI reference architecture. En J. D. Harguess, N. D. Bastian, & T. L. Pace (Eds.), *Assurance and Security for AI-enabled Systems* (p. 27). SPIE. https://doi.org/10.1117/12.3027227

Vasquez, D. A. (2025). *Análisis comparativo de algoritmos de machine learning para predecir morosidad en clientes afiliados a la entidad financiera San Francisco de Mocupe* [Universidad Señor de Sipan]. https://alicia.concytec.gob.pe/vufind/Record/USSS_6feaa58844c20f41f61c7301da9599b5

Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

Wang, H., Mao, K., Wu, W., & Luo, H. (2023). Fintech inputs, non-performing loans risk reduction and bank performance improvement. *International Review of Financial Analysis*, *90*, 102849. https://doi.org/10.1016/j.irfa.2023.102849

Wang, K. S., Harjono, K. J., & Lawrence, R. (2025). Towards Reliable Agents: Benchmarking Customized LLM-Based Retrieval-Augmented Generation Frameworks with Deployment Validation. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 655-661. https://doi.org/10.18653/v1/2025.naacl-industry.53

Wang, L., & Zhang, W. (2023). A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from Chinese manufacturing companies. *Information Processing & Management*, *60*(3), 103267. https://doi.org/10.1016/j.ipm.2023.103267

Wang, L.-F., Zhu, H.-J., Wang, C., Yan, F., & Qu, C.-Z. (2025). Logistic regression-based risk prediction of aortic adverse remodeling following thoracic endovascular aortic repair in patients with aortic dissection. *World Journal of Cardiology*, *17*(12). https://doi.org/10.4330/wjc.v17.i12.112978

Wong, L. R., Mauricio, D. S., & Rodriguez, G. D. (2017). A systematic literature review about software requirements elicitation. Journal of Engineering Science and Technology, 12(2), 296-317.

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

Wu, Z., Dong, Y., Li, Y., & Shi, B. (2025). Unleashing the power of text for credit default prediction: Comparing human-written and generative AI-refined texts. *European Journal of Operational Research*, *326*(3), 691-706. https://doi.org/10.1016/j.ejor.2025.04.032

Yang, T., & Ying, Y. (2023). AUC Maximization in the Era of Big Data and AI: A Survey. *ACM Computing Surveys*, *55*(8), 1-37. https://doi.org/10.1145/3554729

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models* (Versión 3). arXiv. https://doi.org/10.48550/ARXIV.2210.03629

Yepmo, V., Smits, G., Lesot, M.-J., & Pivert, O. (2024). Leveraging an Isolation Forest to Anomaly Detection and Data Clustering. *Data & Knowledge Engineering*, *151*, 102302. https://doi.org/10.1016/j.datak.2024.102302

Yin, Z., Hu, B., & Chen, S. (2024). *Predicting Employee Turnover in the Financial Company: A Comparative Study of CatBoost and XGBoost Models*. Computer Science and Mathematics. https://doi.org/10.20944/preprints202410.0072.v1

Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S., & Meng, S. (2024). Advanced User Credit Risk Prediction Model Using LightGBM, XGBoost and Tabnet with SMOTEENN. *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 876-883. https://doi.org/10.1109/ICPICS62053.2024.10796247

Zavala, D. M. K., & Poma, J. A. (2024). *Implementación de un Dashboard inteligente para el análisis predictivo del riesgo crediticio en portafolios de inversión de un holding financiero* [Universidad Peruana de Ciencias Aplicadas]. https://repositorioacademico.upc.edu.pe/entities/publication/e64fcb6a-20e8-5833-a6c0-f58ac58e8338

Zegarra Valdivia, D., & Gil Aguilar, G. (2025). Elaboración de perfiles y el scoring crediticio: Caracterización jurídica y aplicación del régimen legal peruano de protección de datos personales. *IUS ET VERITAS*, (71), 13-25. https://doi.org/10.18800/iusetveritas.202502.001

Zhang, Z., Chen, R., & Stadie, B. C. (2026). *All Leaks Count, Some Count More: Interpretable Temporal Contamination Detection and Mitigation in LLM Backtesting* (Versión 2). arXiv. https://doi.org/10.48550/ARXIV.2602.17234

Zou, Y., & Gao, C. (2022). Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring. *Algorithms*, *15*(5), 149. https://doi.org/10.3390/a15050149

ANEXOS

ANEXO A. Documento de Analisis de Software

Especificación de Requisitos Funcionales

Fase 1: Motor Analítico

Requerimiento funcional N°:	Nombre
RQF01	Cálculo de Probabilidad y Score Crediticio
Tipo:	Prioridad:
Backend / Modelo Predictivo	Alta
Descripción:	
El modelo debe calcular la probabilidad de impago de un cliente y aplicar una transformación para normalizar el resultado en una escala de score de 0 a 1000.	

Requerimiento funcional N°:	Nombre
RQF02	Cálculo de Explicabilidad Nativa (SHAP)
Tipo:	Prioridad:
Backend / Modelo Predictivo	Alta
Descripción:	
El sistema debe calcular los Valores de Shapley (SHAP) por cada predicción, determinando el peso exacto de cada variable predictora para justificar el nivel de riesgo.	

Requerimiento funcional N°:	Nombre
RQF03	Automatización de Inferencia en la Nube
Tipo:	Prioridad:
Backend / Arquitectura Cloud	Alta
Descripción:	
El modelo debe ejecutarse automáticamente de manera mensual mediante un pipeline, procesando la cartera activa y persistiendo los resultados (probabilidad, score y SHAP) en la infraestructura de la nube.	

Fase 2: Plataforma de Seguimiento

Requerimiento funcional N°:	Nombre
RQF04	Registro y Aprobación de Accesos
Tipo:	Prioridad:
Seguridad / Gestión de Identidades	Alta
Descripción:	
El sistema debe permitir el registro de nuevos usuarios, manteniéndolos en estado inactivo hasta que un usuario Administrador apruebe su ingreso a la plataforma.	

Requerimiento funcional N°:	Nombre
RQF05	Personalización de Perfil de Usuario
Tipo:	Prioridad:
Frontend / Interfaz de Usuario	Baja
Descripción:	
El usuario debe poder actualizar su información personal (Nombre, foto, perfil), almacenando los cambios en memoria persistente.	

Requerimiento funcional N°:	Nombre
RQF06	Gestión de Roles y Permisos Dinámicos
Tipo:	Prioridad:
Seguridad / Frontend	Alta
Descripción:	
El sistema debe tener roles específicos (Analista, Validador, Monitoring). Las vistas con la cual interactuará el usuario deben restringirse dinámicamente según los permisos del rol activo.	

Requerimiento funcional N°:	Nombre
RQF07	Monitoreo de Performance y Rendimiento
Tipo:	Prioridad:
Backend / Analítica de Datos	Alta
Descripción:	
El sistema debe mostrar el evolutivo mensual de métricas de rendimiento, incluyendo Gini, KS y Matriz de Confusión, en tableros interactivos.	

Requerimiento funcional N°:	Nombre
RQF08	Monitoreo de Estabilidad y Data Drift
Tipo:	Prioridad:
Backend / Analítica de Datos	Alta
Descripción:	
El sistema debe mostrar métricas de estabilidad, como el PSI del modelo general y por variables y debe generar una alerta visual si el valor del PSI supera los umbrales de riesgos definidos. Además se debe mostrar ratios, cantidades para cada variable.	

Requerimiento funcional N°:	Nombre
RQF09	Gestión de Chats y Persistencia
Tipo:	Prioridad:
Frontend / Base de Datos	Media
Descripción:	
La plataforma debe permitir al usuario crear, nombrar, editar y eliminar ventanas de chat interactivo con un Agente LLM, manteniendo un registro persistente de las conversaciones pasadas.	

Requerimiento funcional N°:	Nombre
RQF10	Consultas Analíticas del Modelo mediante LLM
Tipo:	Prioridad:
Inteligencia Artificial / LangChain	Alta
Descripción:	
El Agente MRM, basado en un LLM, debe interpretar texto en lenguaje natural para responder preguntas de performance, estabilidad, explicabilidad y detalle de usuarios.	

Requerimiento funcional N°:	Nombre
RQF11	Consultas Normativas (RAG)
Tipo:	Prioridad:
Inteligencia Artificial / RAG	Alta
Descripción:	
El Agente debe consultar documentos normativos de la SBS y manuales metodológicos para responder consultas basadas 100% en esta referencia y evitar alucinaciones.	

Especificación de Requisitos No Funcionales

Código	Nombre del Requerimiento	Característica ISO 25010	Descripción
RNF01	Eficiencia en Procesamiento Masivo	Eficiencia de Desempeño	Capacidad de procesar millones de registros históricos en tiempos operativos viables.
RNF02	Cifrado y Control de Accesos	Seguridad	Todo el tráfico debe estar cifrado y el acceso a los microservicios debe validarse exclusivamente mediante tokens.
RNF03	Alta Disponibilidad y Tolerancia a Fallos	Fiabilidad	El sistema debe mantener la continuidad del servicio
RNF04	Navegación Web Dinámica	Usabilidad	La interfaz debe operar garantizando transiciones fluidas entre tableros sin recargar el navegador.
RNF05	Arquitectura Desacoplada y Modular	Mantenibilidad	Separación estricta entre la capa de presentación, lógica de negocio y procesamiento de IA permitiendo actualizaciones independientes.

Historias de Usuario (Product Backlog)

Historia de Usuario #1	
Cálculo de Probabilidad y Score	
Como	Analista de Riesgo, Analista de Producto, Data Science, Model Risk Monitoring, Model Risk Validator
Quiero	Tener la probabilidad de impago en una escala de 0 a 1000
Para Poder	Realizar análisis de provisiones y riesgo de cartera

Como usuario quiero tener la probabilidad de impago en una escala de 0 a 1000, para poder realizar análisis de provisiones y riesgo de cartera.	
Actividad	Story Points (S.P.)
Programación del pipeline de inferencia.	2
Integración de fórmula de transformación.	0.5
Pruebas Unitarias y Testing	1
Pase a producción y documentación	2
Estimación Total	5.5 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• La data debe estar disponible cada mes que se actualiza el stock de clientes • La data debe estar disponible en una base de datos • Los datos de los clientes deben estar encriptados • Se debe visualizar el score, la probabilidad y las variables de los clientes

Historia de Usuario #2	
Explicabilidad Nativa SHAP	
Como	Analista de Riesgo, Analista de Producto, Data Science, Model Risk Monitoring, Model Risk Validator
Quiero	Obtener el peso de cada variable predictora en la inferencia del modelo
Para Poder	Comprender por qué un cliente fue calificado con alto o bajo riesgo

Como usuario quiero el peso individual de cada variable predictora, para poder Comprender por qué un cliente fue calificado con alto o bajo riesgo.	
Actividad	Story Points (S.P.)
Cálculo de Valores SHAP por cliente	1
Persistencia de aportes en la base de datos	1
Programación de Pipeline	1
Pase a producción y documentación	2
Estimación Total	5 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• Los valores de aporte deben estar expresados en log-odds • Los valores deben estar disponibles en una base de datos • Los valores deben estar disponibles para todos los clientes y variables • La data debe estar lista para su consumo luego de la inferencia del modelo.

Historia de Usuario #3	
Métricas de Monitoreo de Estabilidad y Performance	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE
Quiero	Obtener las métricas de monitoreo para performance y estabilidad del modelo
Para Poder	Analizar si el modelo tiene buen rendimiento y presenta estabilidad global y a nivel de variables.

Como usuario quiero obtener las métricas de monitoreo para performance y estabilidad del modelo, para analizar si el modelo tiene un buen rendimiento y presenta estabilidad global/variables.	
Actividad	Story Points (S.P.)
Diseño del flujo de cálculo de indicadores	4
Creación del Script de Monitoreo	4
Persistencia de Datos en Nube	2
Automatización del pipeline completo	3
Pase a producción y documentación	2
Estimación Total	15 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• Las metricas de performance y estabilidad se deben almacenar en tablas individuales • El resultado de las metricas deben estar particionadas por mes y modelo. • La tabla debe tener una estructura que facilite su entendimiento.

Historia de Usuario #4	
Registro de nuevo Usuario	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Crear una cuenta en la plataforma Web
Para Poder	Interactuar con el Sistema Web de Seguimiento

Como usuario quiero crear una cuenta en la plataforma Web, para interactuar con el sistema Web de Seguimiento	
Actividad	Story Points (S.P.)
Creación del Formulario de Front-end	1
Creación de pool de usuarios	1
Creación de Bucket para almacenar multimedia	2
Creación de Atributos de pool de usuarios	1
Pase a producción y documentación	2
Estimación Total	7 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• El formulario de registro debe pedir correo y contraseña. • Se debe enviar un mensaje de confirmacion, indicando que falta aprobacion de usuario. • Solo se permite crear un usuario con un solo correo. • La contraseña debe tener al menos 8 caracteres y cumplir con la política de seguridad (incluir números, mayúsculas, minúsculas o símbolos).

Historia de Usuario #5	
Aprobación de nuevo usuario	
Como	Administrador
Quiero	Aprobar el registro nuevo usuarios
Para Poder	Brindar acceso y puedan interactuar con el sistema

Como administrador quiero aprobar el registro de nuevos usuarios para poder brindar acceso y puedan interactuar con el sistema,	
Actividad	Story Points (S.P.)
Configuración de usuario administrador con privilegios elevados	1
Configuración de Pool de usuarios	1
Estimación Total	2 (S.P.)
PRIORIDAD	Media

Criterios de Aceptación
• La aprobacioń se realiza 100% manual mediante la interfaz de la nube • El administrador debe visualizar la lista de usuarios • El administrador puede aproba, rechazar y editar atributos de los usuarios.

Historia de Usuario #6	
Inicio de sesión	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Iniciar Sesión en la plataforma Web
Para Poder	Interactuar con las funcionalidades de acuerdo con mi rol asignado
Como usuario quiero iniciar sesión en la plataforma web para interactuar con las funcionalidades de acuerdo con mi rol asignado.	
Actividad	Story Points (S.P.)
Creación del Formulario de Inicio de sesión	1

Recuperación de Atributos y multimedia	3
Estimación Total	2 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• El formulario debe pedir solo correo y contraseña. • En el campo de contraseña existira un boton para hacer visible y ver si existen errores • Si el usuario no esta aun aprobado se mostrara un pop-up de error • Se mostrara contraseña incorrecta si no coincide • Una vez autenticado por primera vez se pedira cargar una foto y el Rol Asignado

Historia de Usuario #7	
Configuración de Perfil	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Editar la configuración de mi perfil
Para Poder	Personalizar mis datos dentro de la plataforma Web
Como usuario quiero editar la configuración de mi perfil para poder personalizar mis datos dentro de la plataforma web.	
Actividad	Story Points (S.P.)
Creación del Formulario para edición de datos	1
Envió y guardado de nueva información	1
Carga y Recuperación de imagen de perfil	1

Estimación Total	3 (S.P.)
PRIORIDAD	Baja

Criterios de Aceptación
• La edicion debe contemplar solo Nombre Completo, puesto y area/departamento. • Los puestos estan restringidos exclusivamnte a Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo • La imagen que se puede cargar es de maximo 2mb en formato JPG o PNG. • Una vez actualizado los datos se debera refrescar en automatico la pagina para mostrar los nuevos resultados.

Historia de Usuario #8	
Panel de Control General	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Visualizar el panel de control general
Para Poder	Observar las principales métricas del modelo
Como usuario quiero visualizar el panel de control general para observar las principales métricas del modelo	
Actividad	Story Points (S.P.)
Creación del frontend de resumen de métricas	2
Conexión con datos almacenados	1
Creación de botones de solicitud de reentrenamiento y análisis completo	2

Estimación Total	5 (S.P.)
PRIORIDAD	Media

Criterios de Aceptación
• El panel de control general debera tener un filtro para cambiar de modelo en caso exista mas de 2 modelos registrados. • Tiene 2 grandes secciones Performance (Gini del Modelo) y estabilidad de variables (PSI) • Cuenta con 2 botones de acciones de gobernanza, que envian un log a la nube. Esta funcion solo esta habilitada para equipo de Validacion y Monitoreo.

Historia de Usuario #9	
Módulo de Reportes	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Visualizar el módulo de reportes
Para Poder	Observar a detalle las métricas de performance general, por variables y estabilidad.
Como usuario quiero visualizar el módulo de reportes para poder observar a detalle las métricas de performance general, por variables y estabilidad.	
Actividad	Story Points (S.P.)
Creación de APIS que expongan la información de las tablas	3
Draft de maqueta de resultados	2
Front-end de módulo de reportes performance general	3

Front-end de módulo de reportes performance por variables	3
Front-end de módulo de reportes estabilidad	3
Creación de filtros	3
Botones de actualización	1
Estimación Total	19 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación
• El modulo de reportes debe contar con 3 secciones: Performance, performance por variables y estabilidad. • Se debera contar con filtros dinamicos para las 3 pestañas. • Se deberan mostrar metricas generales y particulares para el modelo seleccionado.

Historia de Usuario #10	
Creación de chat Agent MRM	
Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Crear una nueva conversación con un agente
Para Poder	Realizar consultas de métricas del modelo, interpretaciones, recomendaciones y preguntas de usuarios especificas dentro del dominio MRM

Como usuario quiero crear una nueva conversación con un agente para poder realizar consultas de métricas del modelo, interpretaciones, recomendaciones y preguntas de usuarios especificas dentro del dominio MRM	
Actividad	Story Points (S.P.)
Creación de herramientas del agente	4
Orquestación del agente	5
Carga de base de datos Vectorial	5
Despliegue del agente	5
Integración como microservicio	3
Creación de nueva conversación en base de datos	2
Guardado de interacción agente-humano	2
Estimación Total	26 (S.P.)
PRIORIDAD	Alta

Criterios de Aceptación

- El agente debe responder utilisando la data del pipeline de monitoreo.
- El agente debe tener acceso a documentos normativos cargados previamente.
- La conversacion debe ser persistente.
- El dominio del agente se restringue solamente a consultas del modelo en el marco del Model Risk Management.
- El agente debe incorporar medidas de seguridad como filtros internos.

Historia de Usuario #11
Historial y Edición de chats

Como	Data Science, Model Risk Monitoring, Model Risk Validator, MLE, Analista de Producto, Analista de Riesgo
Quiero	Visualizar mi historial de chat y realizar la edición de títulos de conversaciones
Para Poder	Etiquetar de manera más Ágil las conversaciones y tener una mejor orientación.
Como usuario quiero visualizar mi historial de chat y realizar la edición de títulos de conversaciones para poder etiquetar de manera más Ágil las conversaciones y tener una mejor orientación.	
Actividad	Story Points (S.P.)
Recuperación de conversaciones de la base de datos	2
Front-end de edición de títulos de conversaciones	1
Estimación Total	3 (S.P.)
PRIORIDAD	Media
Criterios de Aceptación	
• La lista de conversaciones debe mostrarse como barra lateral dentro de la pestaña del agente. • Las conversaciones permitiran editar solo el titulo.	

ANEXO B. Documento de Arquitectura de software

DOCUMENTO DE ARQUITECTURA DE SOFTWARE

Proyecto	Desarrollo de un Modelo de Machine Learning usando XGBoost para medir la probabilidad de impago de clientes en una entidad bancaria, 2026
Autor	Eduardo Rafael Jáuregui Romero
Versión del DAS	1.0
Fecha	Septiembre de 2026

Control de cambios

Versión	Descripción	Autor	Fecha
1.0	Primera versión del documento de arquitectura.	Eduardo Rafael Jáuregui Romero	20/09/2026

Tabla de contenido

1. Introducción
2. Contexto, objetivos y alcance
3. Requisitos y drivers arquitectónicos
4. Macro arquitectura y modelo de vistas
5. Arquitectura del motor analítico y pipeline del modelo
6. Arquitectura de inferencia, SHAP y persistencia

7. Arquitectura de monitoreo
8. Arquitectura de la plataforma web
9. Arquitectura del Agente MRM y RAG
10. Vista de casos de uso e historias de usuario
11. Vista lógica y de componentes
12. Vista de procesos
13. Vista de datos
14. Vista física y de despliegue
15. Seguridad
16. Atributos de calidad y tácticas
17. Operación, observabilidad y despliegue
18. Conclusiones y anexos

1. Introducción

1.1 Propósito

El propósito del presente Documento de Arquitectura de Software (DAS) es describir, de manera integral, la estructura técnica de la solución desarrollada para estimar la probabilidad de impago de clientes de préstamos personales y operar su ciclo de vida en producción.

La solución no se limita al algoritmo XGBoost, sino que incluye los siguientes frentes:

- Construcción de datos, preprocesamiento, entrenamiento, optimización, inferencia mensual, transformación a score, explicabilidad SHAP,
- Monitoreo de performance y estabilidad,
- Plataforma web Model Risk Management Suite, control de acceso y un agente LLM con herramientas analíticas y RAG.

1.2 Audiencia

El documento está dirigido a developers, Data Science, Machine Learning Engineering (MLES), Model Risk Monitoring (MRM), Model Risk Validator (MRV), analistas de riesgo/producto, arquitectura y auditoría.

Su objetivo es permitir que un tercero pueda entender qué componentes existen, cómo interactúan y qué requisitos justifican cada decisión.

1.3 Alcance

- Motor analytics de probabilidad de impago y score crediticio.
- Pipeline de preparación, entrenamiento, optimización y evaluación del modelo.
- Pipeline mensual de inferencia y cálculo de SHAP.
- Pipeline de monitoreo de performance y estabilidad.
- Plataforma web SPA y backend serverless.
- Gestión de identidad, roles, permisos y perfil.
- Agente MRM con LangChain, patrón ReAct, tools analíticas y RAG.
- Persistencia, consumo y exposición de datos analíticos.
- Historias de usuario, casos de uso, atributos de calidad y trazabilidad.

1.4 Fuera de alcance / no especificado

Por temas de confidencialidad no se detallará los esquemas físicos completos de todas las tablas de negocio. Por ello, este DAS representa esos elementos como componentes lógicos de persistencia sin indicar unas reglas/tecnologías exactas. Tampoco se diseña aquí el core transaccional bancario, se asume que la solución consume información disponible para el modelo y sus procesos de seguimiento.

1.5 Definiciones y abreviaturas

Sigla	Definición
DAS	Documento de Arquitectura de Software
MRM	Model Risk Management
MLE	Machine Learning
OOT	Out of Time
SHAP	Shapley Additive exPlanations
PSI	Population Stability Index
SPA	Single Page Application
JWT	JSON Web Token
RAG	Retrieval-Augmented Generation
ReAct	Reasoning + Acting
FaaS	Function as a Service
S3	Amazon Simple Storage Service

ECS	Amazon Elastic Container Service
ECR	Amazon Elastic Container Registry
HU	Historia de Usuario
RQF	Requerimiento funcional
RNF	Requerimiento no funcional

2. Contexto, objetivos y alcance arquitectónico

2.1 Contexto del sistema

La solución soporta el ciclo completo de un modelo de riesgo de crédito para préstamos personales (PP). El estudio parte de 1,735,340 clientes y 582 variables, reducidas a 48 mediante filtros de calidad, análisis bivariado y WOE. El modelo final XGBoost fue seleccionado y optimizado en cloud (AWS), y luego integrado con mecanismos de explicabilidad, monitoreo y consumo web.

2.2 Objetivos arquitectónicos

- Automatizar el procesamiento mensual de los grandes volúmenes de información (cerca de 1 millon de registros por mes).
- Lograr aislar la capa de presentación de la lógica analítica y del procesamiento de IA.
- Garantizar la trazabilidad desde requerimientos e historias de usuario hasta componentes técnicos finales.
- Proteger datos y servicios mediante autenticación, tokens y comunicaciones cifradas en cloud.
- Facilitar observabilidad del modelo mediante performance, estabilidad y explicabilidad.
- Permitir consultas analíticas y normativas en lenguaje natural sin mezclar responsabilidades con el backend serverless.

2.3 Principios de diseño

Principio	Aplicación
Responsabilidad única	Lambdas orientadas solo a un tipo de lógica. Se cuenta con módulos separados para performance, estabilidad e IA (Agente).
Desacoplamiento	Uso de APIS y microservicios independientes, frontend separado,
Serverless donde aplica	Backend de APIs y monitoreo basado en servicios a demanda.

Procesamiento especializado	Fargate para despliegue del agente IA, sageMaker Processing Job para optimización del modelo. Athena para almacenamiento de datos en lake.
Persistencia analítica consultable	Resultados estructurados en S3 y consumibles mediante Athena.
Seguridad por identidad	Cognito, JWT, roles y credenciales temporales para S3.
Explicabilidad por diseño	SHAP calculado como salida para interpretación de inferencia de cara a negocio.

3. Requisitos y drivers arquitectónicos

3.1 Requisitos funcionales

Código	Nombre	Tipo	Prioridad	Resumen
RQF01	Cálculo de Probabilidad y Score Crediticio	Backend / Modelo	Alta	Calcular probabilidad de impago y normalizar a score 0–1000.
RQF02	Explicabilidad Nativa (SHAP)	Backend / Modelo	Alta	Calcular SHAP por predicción y variable.
RQF03	Automatización de Inferencia en la Nube	Cloud	Alta	Ejecutar mensualmente inferencia y persistir probabilidad, score y SHAP.
RQF04	Registro y Aprobación de Accesos	Identidad	Alta	Registrar usuarios inactivos hasta aprobación administrativa.
RQF05	Personalización de Perfil	Frontend	Baja	Actualizar nombre, foto y perfil en almacenamiento persistente.
RQF06	Roles y Permisos Dinámicos	Seguridad / Frontend	Alta	Restringir vistas según rol activo.
RQF07	Monitoreo de Performance	Analítica	Alta	Mostrar Gini, KS y matriz de confusión por mes.
RQF08	Monitoreo de Estabilidad / Drift	Analítica	Alta	Mostrar PSI, ratios, cantidades y alertas visuales.

RQF09	Gestión de Chats y Persistencia	Frontend / Datos	Media	Crear, nombrar, editar y eliminar chats con historial persistente.
RQF10	Consultas Analíticas mediante LLM	IA / LangChain	Alta	Responder consultas de performance, estabilidad, SHAP y clientes.
RQF11	Consultas Normativas (RAG)	IA / RAG	Alta	Responder usando documentos normativos/metodológicos recuperados.

3.2 Requisitos no funcionales

Código	Nombre	ISO 25010	Descripción
RNF01	Eficiencia en procesamiento masivo	Eficiencia de desempeño	Procesar millones de registros en tiempos operativos viables.
RNF02	Cifrado y control de accesos	Seguridad	Tráfico cifrado y acceso a microservicios validado mediante tokens.
RNF03	Alta disponibilidad y tolerancia a fallos	Fiabilidad	Mantener continuidad del servicio.
RNF04	Navegación web dinámica	Usabilidad	Transiciones entre tableros sin recarga del navegador.
RNF05	Arquitectura desacoplada y modular	Mantenibilidad	Separar presentación, negocio y procesamiento de IA.

3.3 Drivers prioritarios

Los drivers que más condicionan la arquitectura son los siguientes:

- Procesamiento masivo,
- Seguridad
- Desacoplamiento
- Explicabilidad,

- Monitoreo continuo y escalabilidad.

Estos drivers justifican el uso de procesamiento administrado en cloud (AWS), mediante la separación SPA/backend/agente y almacenamiento analítico consultable.

4. Macroarquitectura y modelo de vistas

Para esta sección se adopta una documentación combinada, para describir, visualizar y documentar la arquitectura se usa el enfoque 4+1 (Kruchten) para organizar escenarios y vistas, complementado con macro arquitectura, atributos de calidad y vistas física/lógica/procesos similares a las plantillas de referencia, tal como lo muestra la Figura 1.

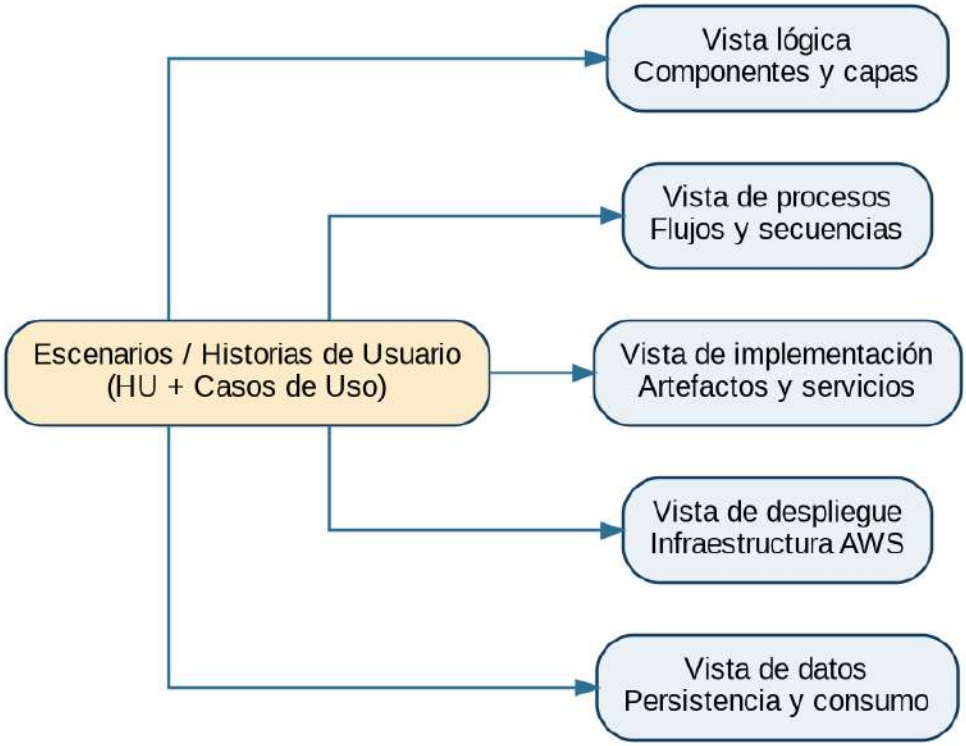


Figura 1. Modelo de vistas arquitectónicas utilizado en el DAS.

La conexión del Modelo – Plataforma web se puede apreciar en la Figura 2 en la cual se muestra el end-to-end del ecosistema de riesgo de modelo

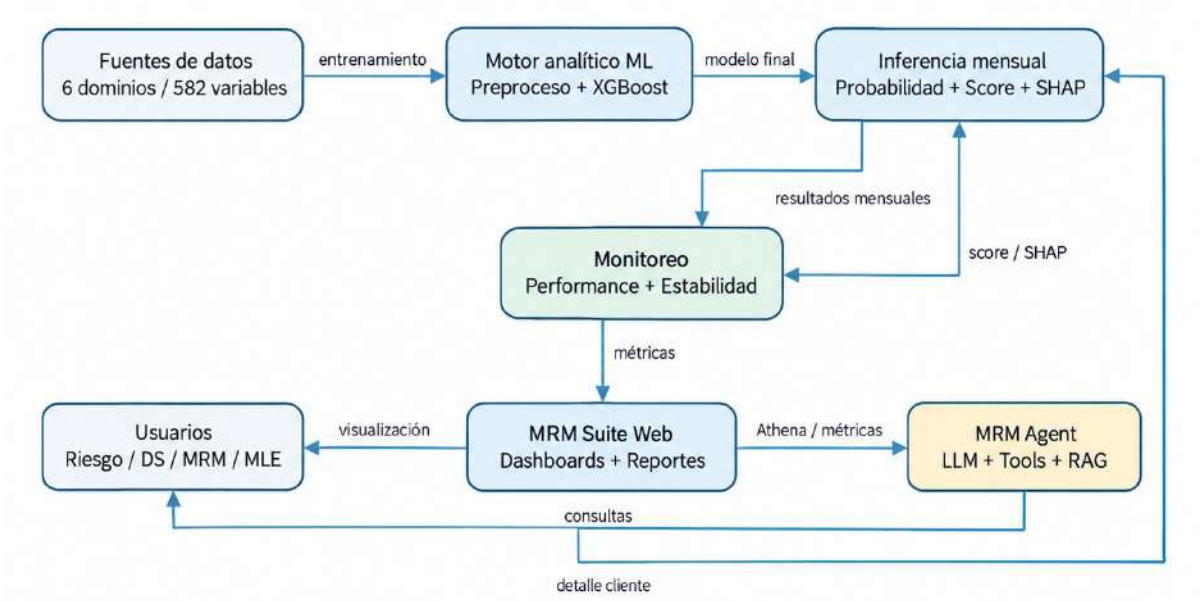


Figura 2. Macroarquitectura end-to-end del ecosistema de riesgo de modelo.

4.1 Dominios arquitectónicos

Dominio	Responsabilidad	Componentes principales
Datos	Consolidar fuentes y resultados analíticos.	Fuentes históricas, S3, Athena, tablas de métricas.
Motor ML	Entrenar, optimizar e inferir probabilidad de impago.	XGBoost, Optuna, SageMaker, objetos de transformación.
Explicabilidad	Justificar predicciones individuales y globales.	SHAP, persistencia de aportes.
Monitoreo	Medir rendimiento y estabilidad en el tiempo.	Step Functions, Glue, S3, Athena.
Plataforma Web	Consumir métricas y administrar usuarios.	SPA, Cognito, API Gateway, Lambda, Amplify, LoadBalancer.
IA conversacional	Responder consultas analíticas y documentales.	LangChain, ReAct, ECS/Fargate, Qdrant, RAG, Langsmith, DynamoDB

5. Arquitectura del motor analítico y pipeline del modelo

5.1 Ciclo de vida del modelo

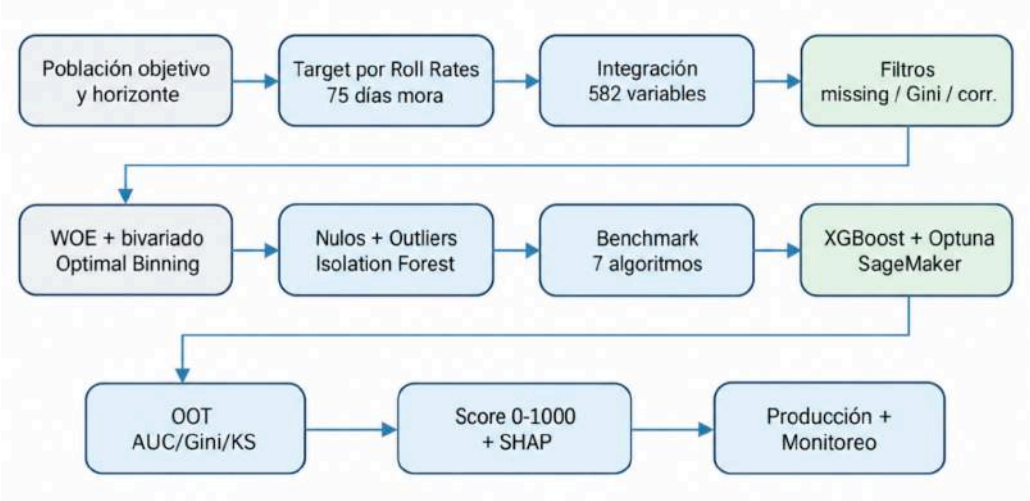


Figura 3. Pipeline integral del ciclo de vida del modelo de probabilidad de impago.

5.2 Construcción del dataset

La arquitectura de datos de entrenamiento integra seis dominios de información: uso crediticio, parque financiero, productos, móvil, sociodemográfico y propiedad. En conjunto conforman 582 variables iniciales. El diseño aplica filtros progresivos hasta 48 variables finales, evitando trasladar ruido y redundancia a las etapas de modelamiento. En la Figura 4 se observa detalle de las fuentes (dominio).

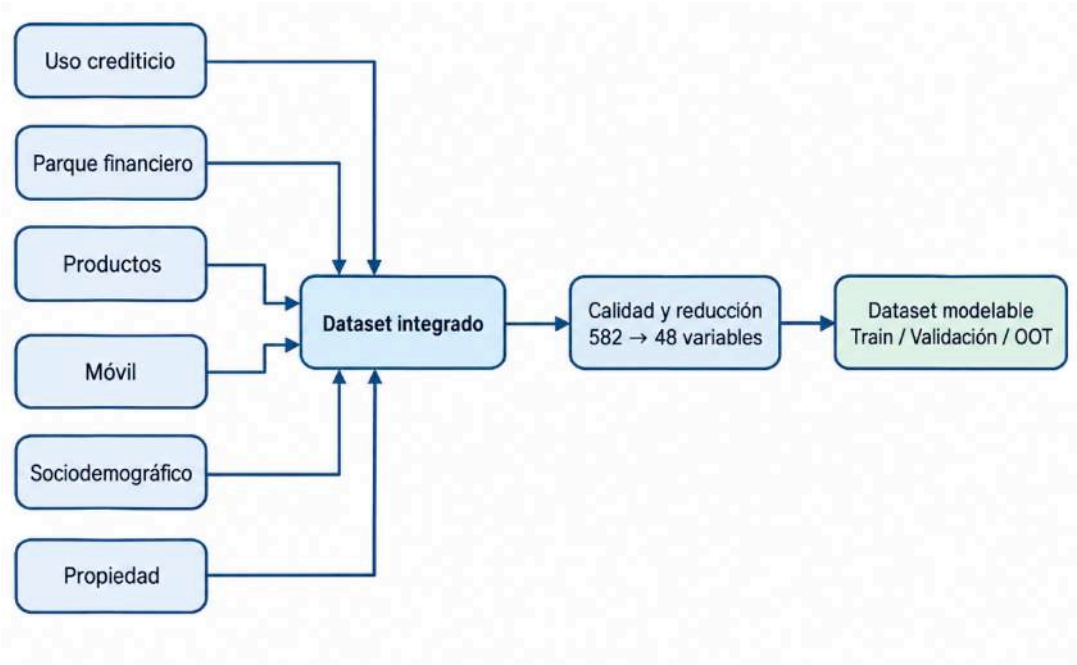


Figura 4. Integración y reducción del dataset de modelamiento.

5.3 Preprocesamiento

Etapa	Objetivo	Salida
Filtros de missings	Retirar variables con ausencia excesiva.	Conjunto depurado.
Gini mínimo	Eliminar variables sin capacidad discriminante suficiente.	Variables con señal predictiva.
Correlación	Reducir redundancia entre predictores.	Set menos colineal.
Análisis bivariado + WOE	Evaluar relación con target y transformar variables.	Variables transformadas.
Toy Model	Priorizar importancia relativa.	Subconjunto candidato.
Tratamiento de nulos	Aplicar imputación referencial según diseño.	Dataset consistente.
Isolation Forest	Detectar y aislar outliers.	Muestras controladas.
Correlación final	Verificar dataset definitivo.	48 variables finalistas.

5.4 Benchmark y selección de algoritmo

El proyecto evalúa Regresión Logística, Árbol de Decisión, Random Forest, AdaBoost, CatBoost, LightGBM y XGBoost. La selección final de XGBoost considera rendimiento, regularización y tratamiento del desbalance. El problema de sobreajuste observado en la comparación inicial se aborda mediante optimización de hiperparámetros y scale_pos_weight.

5.5 Optimización en Amazon SageMaker

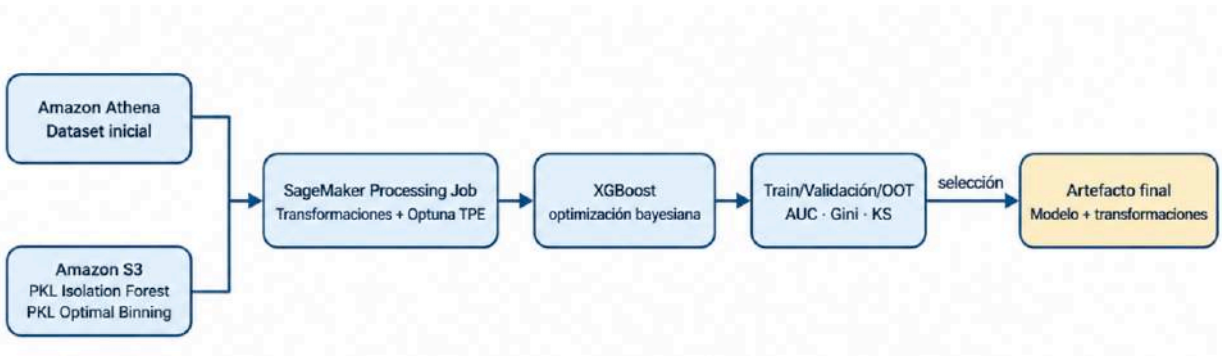


Figura 5. Arquitectura cloud de optimización de hiperparámetros en Amazon SageMaker.

El Processing Job consulta el dataset desde Athena y recupera desde S3 los objetos PKL requeridos por las transformaciones, particularmente Isolation Forest y Optimal Binning. Optuna ejecuta optimización bayesiana bajo TPE sobre XGBoost. El entrenamiento utiliza

periodos 202106–202208 y un conjunto OOT 202209–202210. El desbalance se controla con scale_pos_weight y parámetros de regularización gamma, lambda y alpha.

5.6 Criterios de aceptación técnicos del motor analítico

- Reproducibilidad del dataset a partir de fuentes y periodos definidos (semilla definida)
- Separación temporal de OOT para evaluación fuera de tiempo.
- Persistencia de objetos de transformación necesarios para inferencia.
- Registro de parámetros e hiperparámetros del modelo final.
- Generación de métricas AUC, Gini y KS para comparación y aprobación.
- Disponibilidad del artefacto final para inferencia mensual (.pkl).

6. Arquitectura de inferencia, score y SHAP

6.1 Pipeline mensual de inferencia



Figura 6. Flujo lógico de inferencia mensual, score y explicabilidad.

El RQF03 establece ejecución automática mensual sobre la cartera activa. El pipeline genera tres familias de salida por cliente: probabilidad de impago, score crediticio y valores SHAP. Todos estos procesos son orquestados mediante el servicio de AWS Step Function, que internamente usa scripts de python/GLUE.

6.2 Transformación a score

La salida probabilística p se transforma mediante Score = (1 − p) × 1000. Valores cercanos a 1000 representan menor riesgo y valores cercanos a 0, mayor riesgo. Esta transformación mejora la interpretabilidad operativa sin alterar la estimación probabilística original.

6.3 Explicabilidad SHAP

Por cada predicción se calculan aportes SHAP por variable a nivel de cliente, a través de un script que se encuentra orquestado. Los criterios del backlog exigen que los aportes estén

expresados en log-odds, disponibles para todos los clientes y variables, persistidos y listos para consumo luego de la inferencia. Esta salida habilita explicaciones individuales y análisis globales de importancia.

6.4 Contrato lógico de salida

Entidad lógica	Campos mínimos soportados por la tesis	Consumidores
Resultado de inferencia	Cliente, periodo, probabilidad, score.	MRM Suite, analistas, agente (Tools).
Explicabilidad	Cliente, periodo, variable, valor SHAP (log-odds).	Reportes, análisis individual, agente.
Variables de cliente	Cliente, periodo, variables explicativas.	Agente y análisis de casos.

7. Arquitectura de monitoreo

7.1 Objetivo

El monitoreo comprueba desempeño real, detecta anomalías estadísticas y valida que la población no cambie estructuralmente respecto del diseño original. La arquitectura documentada en la Figura 7 detalla a alto nivel el pipeline desplegado en AWS mediante Step Function.

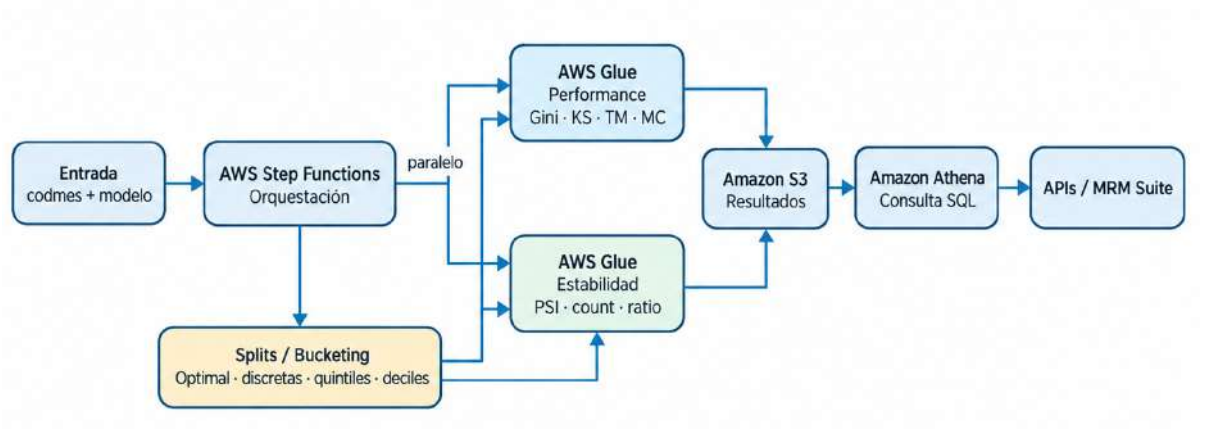


Figura 7. Arquitectura serverless para cálculo de performance y estabilidad.

7.2 Orquestación

AWS Step Functions recibe el mes de proceso (codmes) y el identificador del modelo. A partir de estos parámetros coordina en paralelo dos procesos AWS Glue: Performance y Estabilidad.

Esta separación permite mantener responsabilidades independientes y habilita evolución de cada cálculo sin conectarlo a la interfaz web, haciendo que el módulo sea una caja separada de las demás funciones.

7.3 Módulo de Performance

- Gini mensual y acumulado.
- KS.
- Tasa de malos (bad rate).
- Matriz de confusión.
- Cálculos a nivel de score y, donde corresponde, de variables.

7.4 Módulo de Estabilidad

- Conteo de registros.
- Ratios de distribución.
- PSI mensual y acumulado.
- Estabilidad global y a nivel de variables.

7.5 Bucketing y persistencia

El cálculo consume puntos de corte auxiliares almacenados en tablas de Athena. Se priorizan cortes obtenidos con Optimal Binning. Para las variables discretas se conservan agrupaciones originales. En ausencia de cortes óptimos se utilizan quintiles. El score se trabaja en deciles. Los resultados se almacenan en S3 y se consultan mediante Athena.

7.6 Salidas y alertas

La plataforma expone métricas de performance, performance por variables y estabilidad. RQF08 exige alertas visuales cuando el PSI supera umbrales de riesgo definidos. Estas alertas no son fijas, si no que se personalizan según lo especificado dentro de la plataforma web.

8. Arquitectura de la plataforma web Model Risk Management Suite

8.1 Estilo arquitectónico

La solución web utiliza una arquitectura Cliente – Servidor moderna, teniendo un cliente SPA desacoplado de un backend serverless. La comunicación se realiza mediante APIs. Esta decisión permite navegación sin recarga completa, escalado independiente del backend y separación clara entre presentación, servicios analíticos y agente de IA. En la Figura 8 se puede ver el diagrama de la arquitectura general del sistema.

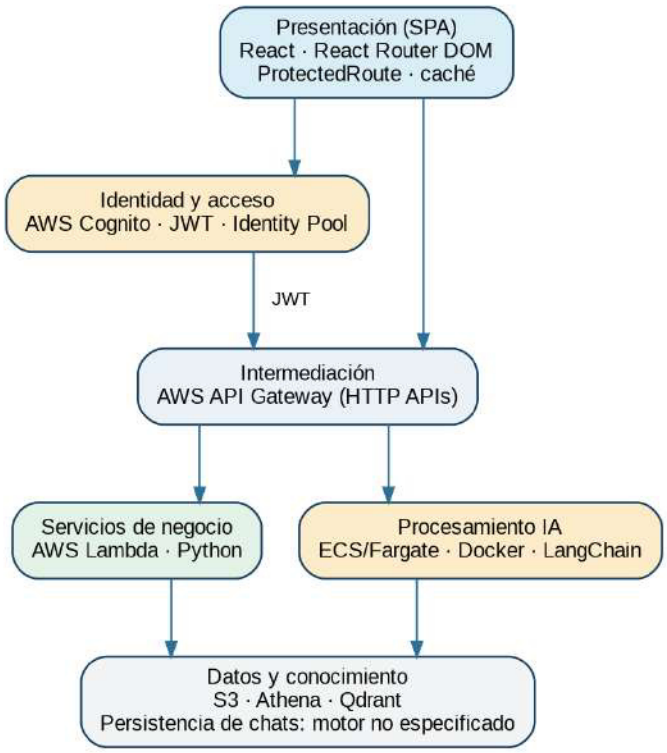


Figura 8. Arquitectura lógica por capas de Model Risk Management Suite.

8.2 Frontend SPA

- React Router DOM para enrutamiento.
- ProtectedRoute para control dinámico de acceso.
- Validación de atributos custom:role y custom:department en Cognito.
- Redirección a /settings cuando el perfil no está completo.
- Caché Singleton en memoria para respuestas JSON.
- Persistencia temporal en LocalStorage con control de tiempo.
- Control de concurrencia reutilizando la misma promesa para consultas duplicadas.
- Sincronización de estado mediante hooks, actualizando solo cuando cambian atributos.

8.3 Gestión de identidad

AWS Cognito autentica al usuario y devuelve un JWT con permisos y rol. Para recursos como fotografía de perfil, Cognito Identity Pool entrega credenciales temporales, evitando exponer llaves permanentes en el navegador.

8.4 Backend serverless

AWS API Gateway recibe peticiones REST, valida tokens, aplica capacidades de enrutamiento, SSL, rate limiting y CORS, y dirige la solicitud hacia Lambdas de negocio o hacia el servicio del agente. Las Lambdas están implementadas en Python y se separan por tipo de dato lógico, siguiendo responsabilidad única.

8.5 Panel general y reportes

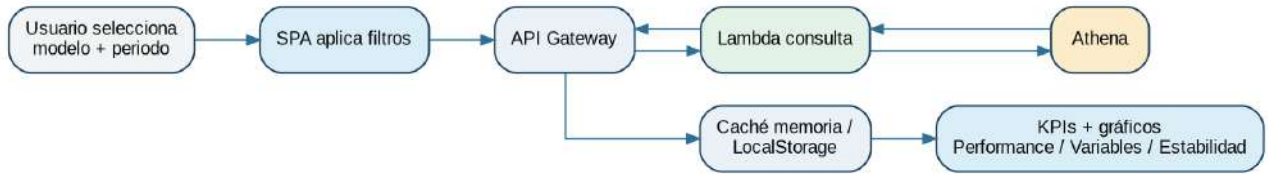


Figura 9. Flujo de consulta del módulo de reportes y visualización.

El panel general permite seleccionar modelos, revisar Gini y PSI y ejecutar acciones de gobernanza restringidas a perfiles de Validación y Monitoreo. El módulo de reportes contiene tres secciones: performance general, performance por variables y estabilidad, con filtros dinámicos y métricas específicas del modelo.



Figura 10. Flujo de acciones de gobernanza desde el panel general.

9. Arquitectura del Agente Model Risk Management

9.1 Objetivo y patrón

El MRM Agent se implementa con el framework de LangChain y bajo el patrón ReAct. El agente interpreta la intención, selecciona una herramienta, observa el resultado y repite el ciclo hasta producir una respuesta dentro del dominio de Model Risk Management.

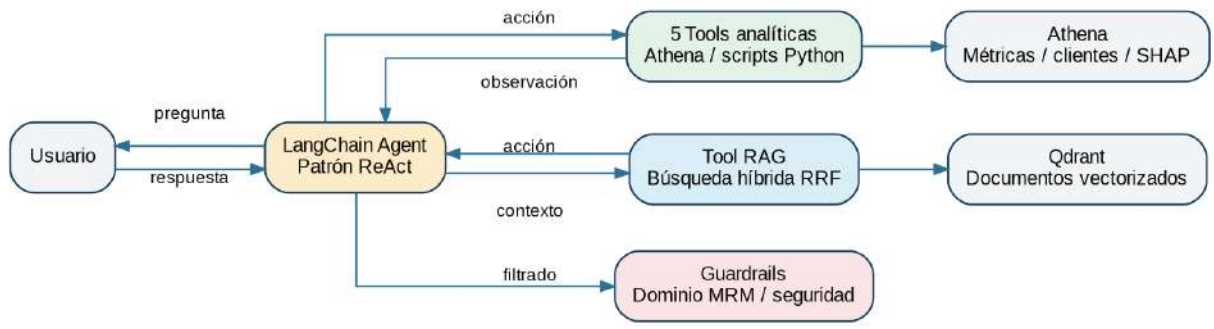


Figura 11. Arquitectura lógica del Agente MRM basado en ReAct.

9.2 Herramientas analíticas documentadas

Tool	Propósito
consultar_metricas_estabilidad	Consultar PSI, count y ratio por mes/variable.
consultar_metricas_performance	Consultar Gini, KS, matriz de confusión, ratio y tasa de malos.
consultar_maestro_cliente	Recuperar perfil, saldos, productos y gestionabilidad de un cliente.
consultar_valores_variables_cliente	Recuperar score y variables explicativas por cliente y periodo.
consultar_valores_shap_cliente	Recuperar aportes SHAP en log-odds y puntos de score.

Las tools consumen scripts Python probados en SageMaker y consultan tablas de Athena. El agente suma una sexta herramienta orientada a RAG para documentación normativa y metodológica.

9.3 RAG

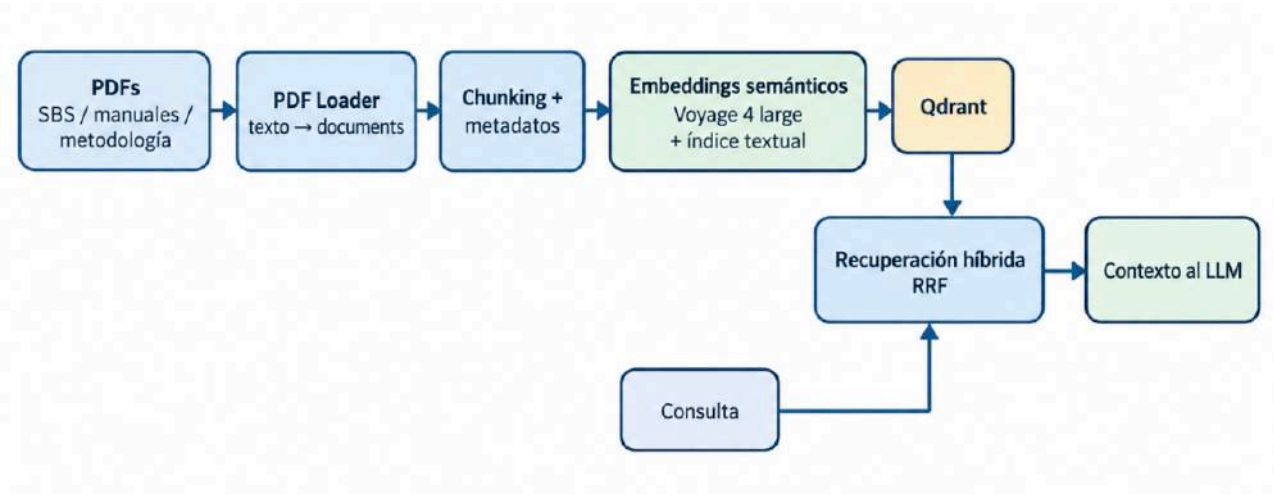


Figura 12. Flujo de ingesta y recuperación RAG.

La fase de ingesta procesa PDFs mediante loader, conversión a documents, chunking y metadatos. Se generan representaciones para búsqueda semántica con Voyage 4 large y búsqueda textual; los vectores se almacenan en Qdrant. En recuperación se usa búsqueda híbrida con RRF para entregar contexto al LLM. La tesis denomina “BM25” al índice textual; este DAS conserva la referencia sin reinterpretarla como otra tecnología.

9.4 Despliegue del agente

El código se conteneriza con Docker, la imagen se almacena en ECR y se ejecuta en ECS usando Fargate Spot. API Gateway expone el servicio mediante HTTPS, API Keys y certificado SSL/TLS. Se incorporan guardrails para restringir el dominio y evitar filtración de datos sensibles o desviaciones fuera de MRM.

10. Vista de casos de uso e historias de usuario

La vista lógica separa responsabilidades funcionales, permitiendo ubicar dónde reside cada decisión y qué dependencias están permitidas.

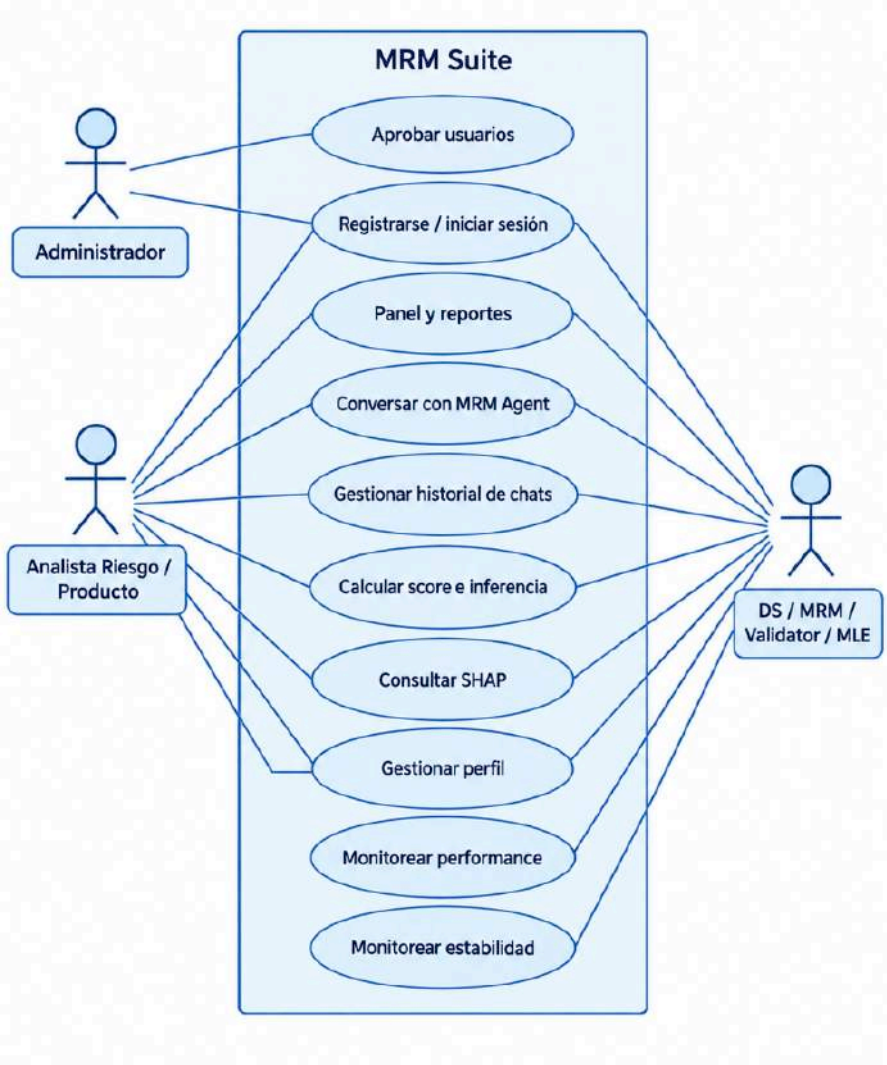


Figura 13. Vista consolidada de actores y casos de uso principales.

10.1 Actores

Actor	Responsabilidades principales

Administrador	Aprobar/rechazar usuarios y editar atributos desde la interfaz de administración disponible.
Data Science	Construcción y análisis del modelo, consulta de métricas y agente.
Model Risk Monitoring	Seguimiento de performance/estabilidad y acciones de gobernanza.
Model Risk Validator	Validación, análisis y acciones de gobernanza.
MLE	Operación técnica de pipelines y plataforma.
Analista de Riesgo	Consumo de score, SHAP, paneles, reportes y agente.
Analista de Producto	Consumo de score, paneles, reportes y agente.

11. Vista lógica y de componentes

La vista lógica separa responsabilidades funcionales, permitiendo ubicar dónde reside cada decisión y qué dependencias están permitidas.

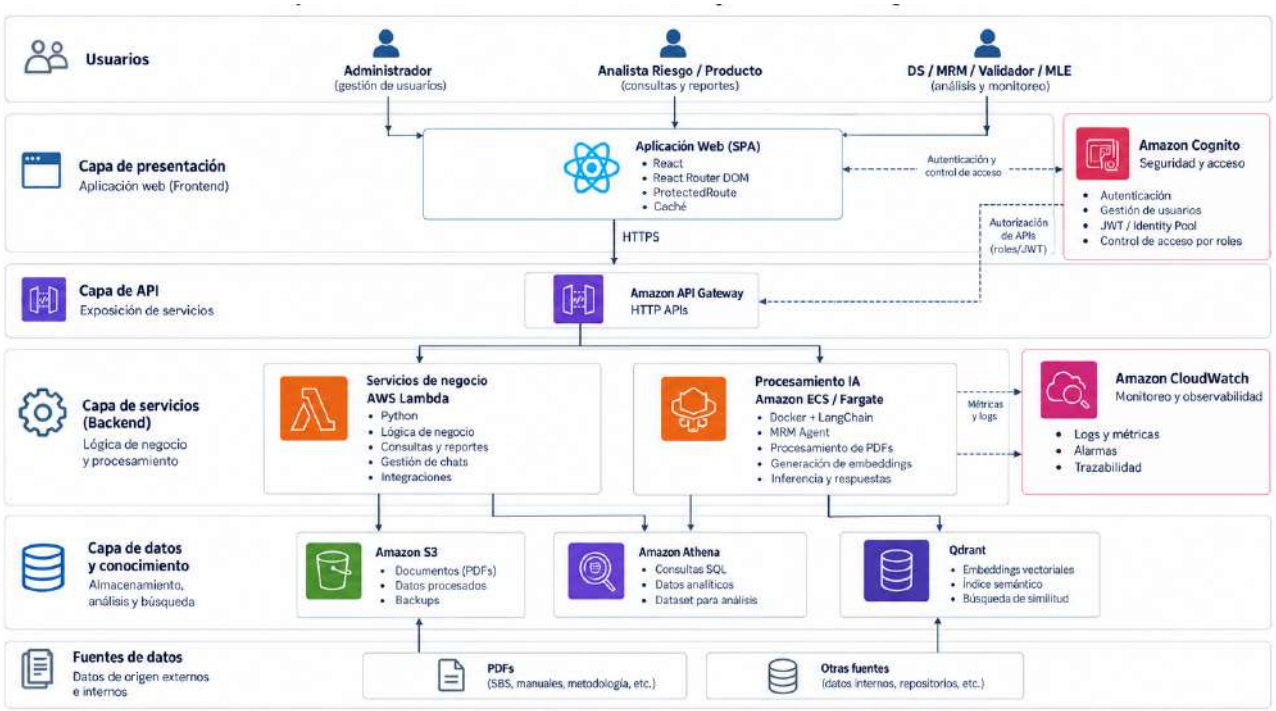


Figura 14. Vista lógica consolidada por capas.

11.1 Catálogo de componentes

Componente	Responsabilidad	Dependencias
SPA	Presentación, navegación, filtros, perfil, chats.	Cognito, API Gateway, S3 temporal.

Cognito	Autenticación, atributos, JWT, credenciales temporales.	SPA, API Gateway, S3.
API Gateway	Frontera HTTP, validación/enrutamiento.	Cognito, Lambda, ECS.
Lambdas	Microservicios de consulta y lógica de APIs.	Athena/S3.
Step Functions	Orquestar monitoreo por mes/modelo.	Glue.
Glue Performance	Calcular Gini/KS/TM/matriz de confusión.	Dataset/splits/S3.
Glue Estabilidad	Calcular PSI/count/ratio.	Dataset/splits/S3.
SageMaker Processing	Optimización y transformaciones de entrenamiento.	Athena, S3, Optuna, XGBoost.
ECS/Fargate	Ejecutar Agente MRM en contenedor.	API Gateway, Athena, Qdrant.
Qdrant	Almacenar/retriever representaciones documentales.	RAG tool.
Athena	Consulta analítica SQL.	S3 y consumidores.
S3	Objetos, resultados, multimedia y artefactos documentados.	SageMaker, Glue, Athena, Cognito Identity Pool.

12. Vista de procesos

12.1 Autenticación y consumo de API

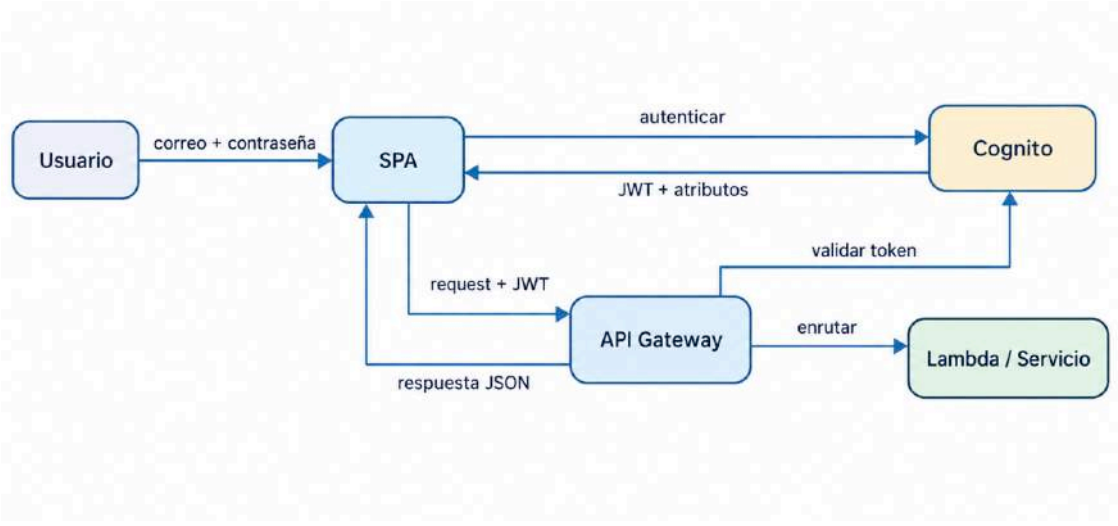


Figura 15. Secuencia lógica de autenticación y consumo de servicios.

12.2 Consulta de reportes

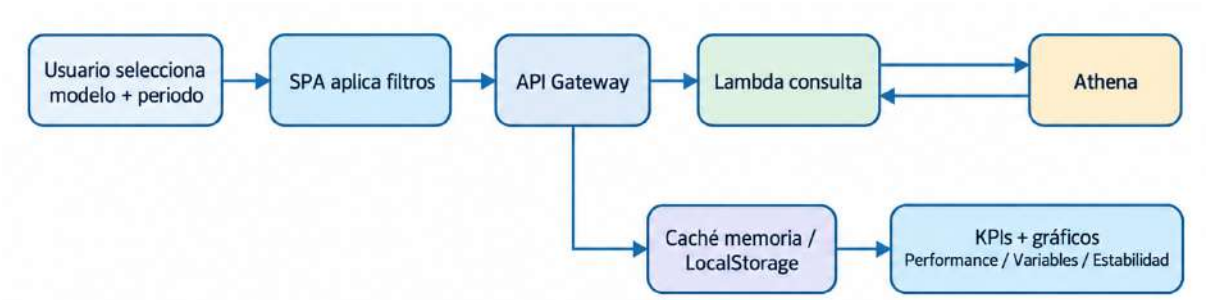


Figura 16. Proceso de consulta y renderizado de métricas.

12.3 Proceso del agente

El proceso conversacional inicia con una pregunta, pasa por controles de dominio/seguridad, utiliza el patrón ReAct para escoger una tool analítica o RAG, recupera observaciones y genera la respuesta. La conversación debe persistir según RQF09 y HU #10/#11; el motor de persistencia no se identifica en la tesis.

12.4 Procesamiento batch

Los procesos analíticos de entrenamiento, inferencia y monitoreo se ejecutan fuera de la interacción síncrona del usuario. Esto evita que el tiempo de cálculo masivo afecte la experiencia de la SPA y permite publicar resultados previamente materializados para consulta.

13. Vista de datos

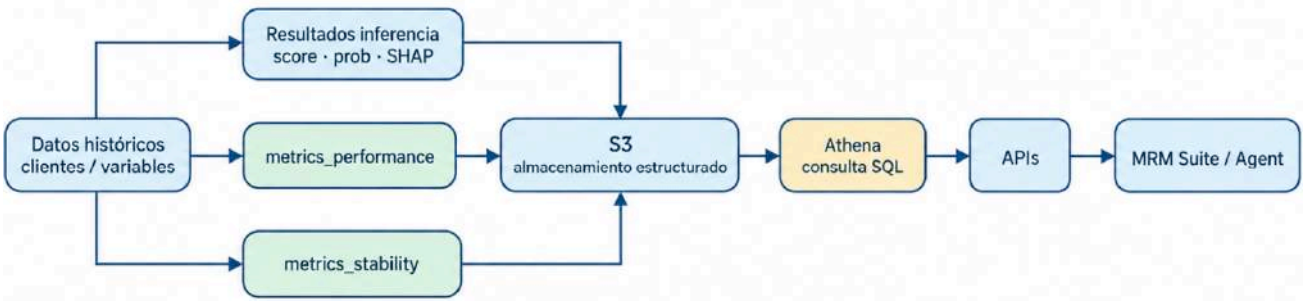


Figura 17. Arquitectura de datos para resultados del modelo y monitoreo.

13.1 Entidades lógicas

Entidad	Granularidad	Uso
Inferencia	cliente × mes × modelo	Probabilidad y score.

SHAP	cliente × mes × modelo × variable	Explicabilidad individual.
Performance	mes × modelo [× variable/tramo]	Gini, KS, TM, matriz de confusión.
Estabilidad	mes × modelo × variable/tramo	PSI, count, ratio.
Splits	modelo × variable × versión	Bucketing y comparabilidad.
Usuario	usuario	Identidad, rol, departamento y perfil.
Conversación	usuario × chat	Historial persistente; tecnología de BD no estructurada DynamoDB
Documento RAG	documento × chunk	Texto, metadatos y representación vectorial en QDrant

13.2 Particionamiento y trazabilidad

El backlog exige que métricas de performance y estabilidad estén particionadas por mes y modelo. Esta clave de partición debe mantenerse como criterio esencial en el almacenamiento y APIs, pues habilita la ejecución, comparación y selección dinámica de modelos en la interfaz.

13.3 Retención y sensibilidad

La tesis exige cifrado y protección de datos de clientes. El DAS establece como criterio de arquitectura que las vistas, APIs y logs expongan únicamente los atributos necesarios para el rol y caso de uso.

14. Vista física y de despliegue

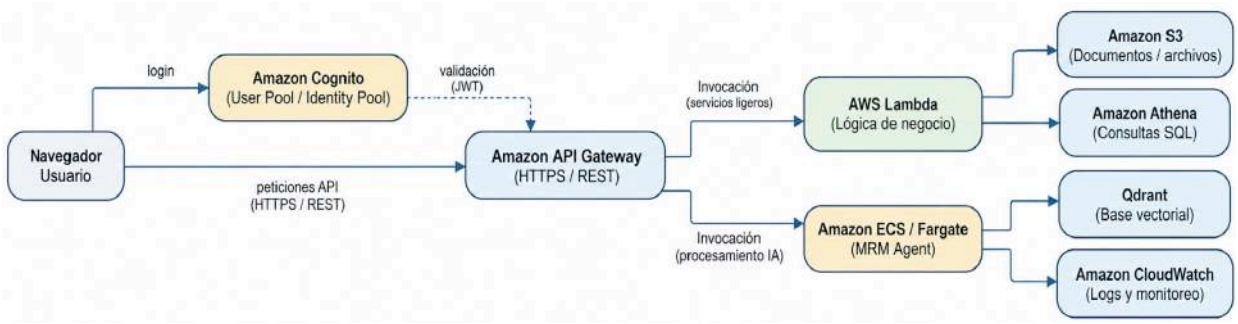


Figura 18. Vista de despliegue de la plataforma y del Agente MRM.

14.1 Nodos y servicios

El usuario opera desde un navegador, en donde AWS Cognito gestiona su identidad, luego API Gateway funciona como puerta de entrada HTTPS. Las lambdas atienden servicios serverless de baja latencia. ECS/Fargate ejecuta el agente contenerizado; ECR almacena la imagen Docker. Athena consulta información analítica y S3 almacena resultados/objetos. Qdrant soporta recuperación hibrida para documental del RAG.

14.2 Separación de cargas

La arquitectura evita ejecutar el LLM dentro de Lambda. El agente reside en ECS/Fargate porque sus tareas son continuas y computacionalmente más intensivas. A su vez, el cálculo batch de monitoreo se ejecuta en Glue y se orquesta con Step Functions, mientras que la optimización de ML usa SageMaker Processing Jobs. Todo el sistema se encuentra separado evitando sobrecargar procesos.

15. Seguridad

15.1 Autenticación y autorización

- AWS Cognito para autenticación.
- JWT para propagar identidad y permisos.
- ProtectedRoute para control de acceso dinámico en SPA.
- Validación de token en API Gateway antes de enrutar a microservicios.
- Credenciales temporales de Cognito Identity Pool para acceso a S3.
- Roles/atributos de usuario para habilitar vistas y acciones de gobernanza.

15.2 Seguridad de transporte y APIs

- HTTPS/SSL-TLS para exposición del agente.
- API Gateway con rate limiting y CORS predefinido.
- API Keys documentadas para acceso al servicio del agente.
- No exponer llaves fijas en el cliente.

15.3 Seguridad del agente

Se incorporan guardrails para restringir el dominio al Model Risk Management, controlar desviaciones y reducir riesgo de filtración de datos sensibles. Los criterios de aceptación exigen filtros internos y uso de fuentes de monitoreo y documentos normativos previamente cargados.

15.4 Riesgos a cubrir en implementación

- Autorización del lado servidor: no depender exclusivamente de la visibilidad de rutas en frontend.
- Evitar registrar datos sensibles completos en logs de API, Glue, Lambda o agente.
- Rotación y administración de API Keys del agente.
- Separación de permisos entre roles técnicos y analistas.
- Validación de tipo/tamaño de archivos de perfil antes de S3.

16. Atributos de calidad y tácticas arquitectónicas

Atributo	Objetivo	Táctica / patrón	Aplicación
Seguridad / Confidencialidad	Evitar acceso no autorizado.	Cognito, JWT, mínimo privilegio, tokens, HTTPS.	Registro/login, APIs, agente.
Eficiencia	Procesar millones de registros.	Glue, SageMaker Processing, Athena, materialización de resultados.	Training/monitoring.
Performance web	Reducir latencia y llamadas repetidas.	Cache Singleton, LocalStorage, control de concurrencia.	SPA.
Disponibilidad	Mantener continuidad.	Servicios gestionados y serverless; desacoplamiento.	Backend y monitoreo.
Mantenibilidad	Cambios aislados.	Capas, microservicios, responsabilidad única.	Toda la solución.
Modificabilidad	Agregar modelos/métricas sin reescribir UI.	Parámetros modelo/periodo, APIs y tablas particionadas.	Plataforma y monitoreo.
Usabilidad	Navegación fluida.	SPA y rutas dinámicas.	Frontend.
Escalabilidad	Ajustar recursos a demanda.	Lambda, Glue, Fargate, SageMaker.	Servicios cloud.

Trazabilidad	Explicar decisiones y ejecuciones.	SHAP, partición por mes/modelo, logs de gobernanza.	Modelo y plataforma.

17. Operación, observabilidad y despliegue

17.1 Frecuencias

Proceso	Frecuencia / disparador	Parámetros
Inferencia	Mensual, tras actualización de stock de clientes.	Periodo y cartera activa.
Monitoreo	Periódico por mes/modelo.	codmes + identificador modelo.
Consultas web	On-demand.	modelo, periodo, filtros.
Agente	On-demand por conversación.	pregunta + contexto de sesión.
RAG ingest	Cuando se incorporan/actualizan documentos.	documentos + metadatos.

17.2 Estrategia de despliegue por componente

- Frontend: artefacto SPA desplegable independientemente del backend.
- Backend: Lambdas y API Gateway versionables de forma desacoplada.
- Agente: imagen Docker → ECR → ECS/Fargate Spot.
- Monitoreo: definición Step Functions + jobs Glue.
- Entrenamiento/optimización: SageMaker Processing Job con dependencias en S3/Athena.

17.3 Observabilidad recomendada como elaboración arquitectónica

La tesis no detalla una plataforma de observabilidad. Como extensión de diseño, se recomienda mantener métricas técnicas por ejecución (inicio/fin, estado, registros procesados, errores) y correlacionarlas por codmes/modelo. Esta recomendación no implica una tecnología específica y debe implementarse con los mecanismos aprobados por la entidad.

18. Conclusiones y anexos

La arquitectura resultante integra en una sola visión el ciclo de vida del modelo y el software que lo operacionaliza. El diseño separa procesamiento batch, servicios interactivos e inteligencia artificial; conserva trazabilidad con los requerimientos y documenta explícitamente dónde la tesis define tecnologías concretas y dónde solo define necesidades lógicas. Esto permite usar el DAS como documento de construcción, operación y validación.

ANEXO C. Documento de Diseño de Interfaz y Experiencia de Usuario (UX/UI)

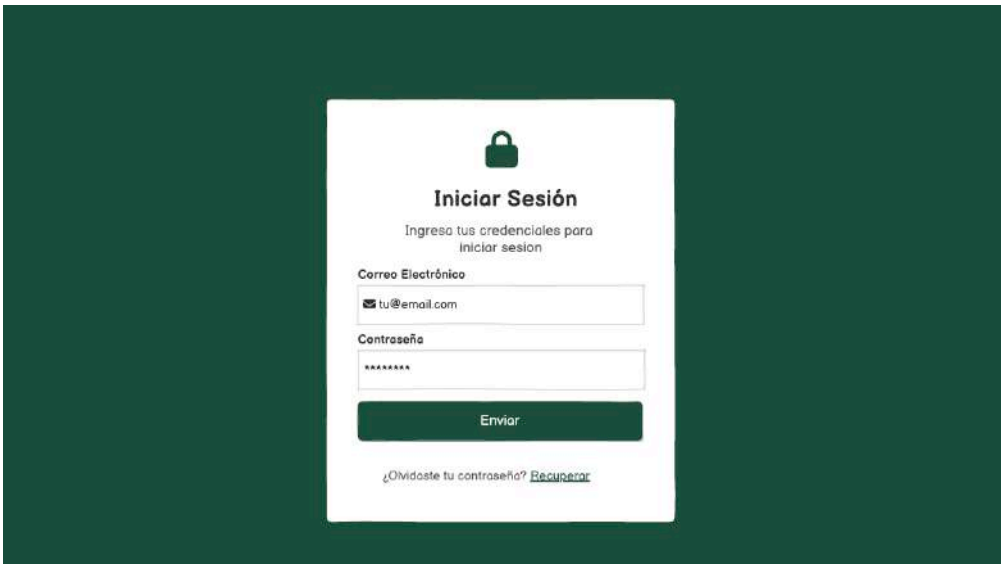
1. Introducción y Objetivos de Diseño

El diseño de la plataforma web "Model Risk Management Suite" tiene un enfoque analítico, interactivo y de reporteria principalmente. Dado que los usuarios finales son analistas, validadores y perfiles técnicos (Data Scentist, MLE, etc.), la interfaz prioriza vistas técnicas, visualización clara de grandes volúmenes de datos (métricas) y un flujo conversacional para interactuar con el Agente MRM (LLM).

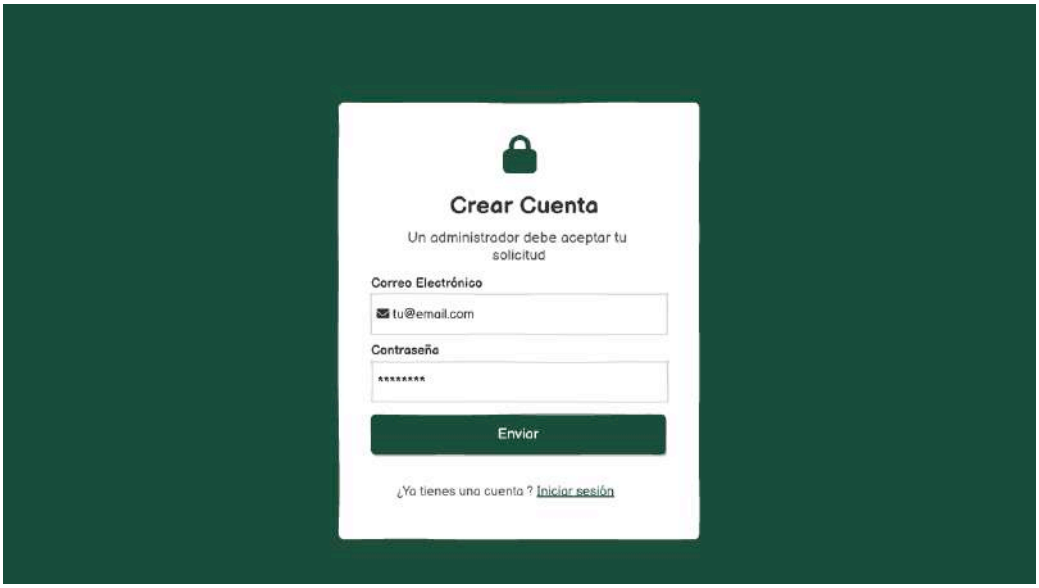
2. Arquitectura de la plataforma (Mapa de navegación)

Para esta seccion se realizara la maquetacion de los wireframes en el software Balsamic. Las agrupaciones son basadas en un flujo normal de un usuario nuevo, las agrupaciones tambien estas relacionando tantos los requisitos funcionales como las historias de usuario.

- **Autenticación**
    - Iniciar Sesión



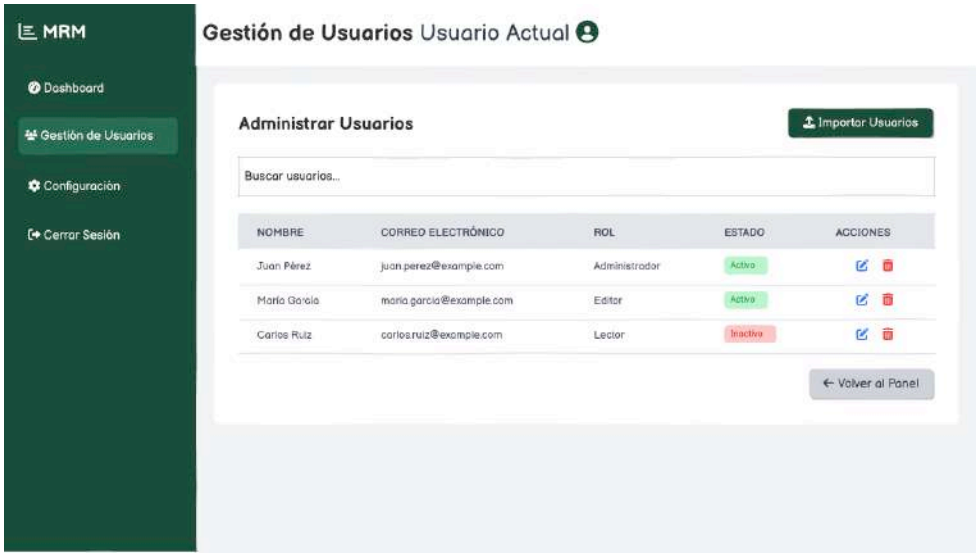
    - Registro



- **Configuración de Usuario**
  - Configuración de perfil.

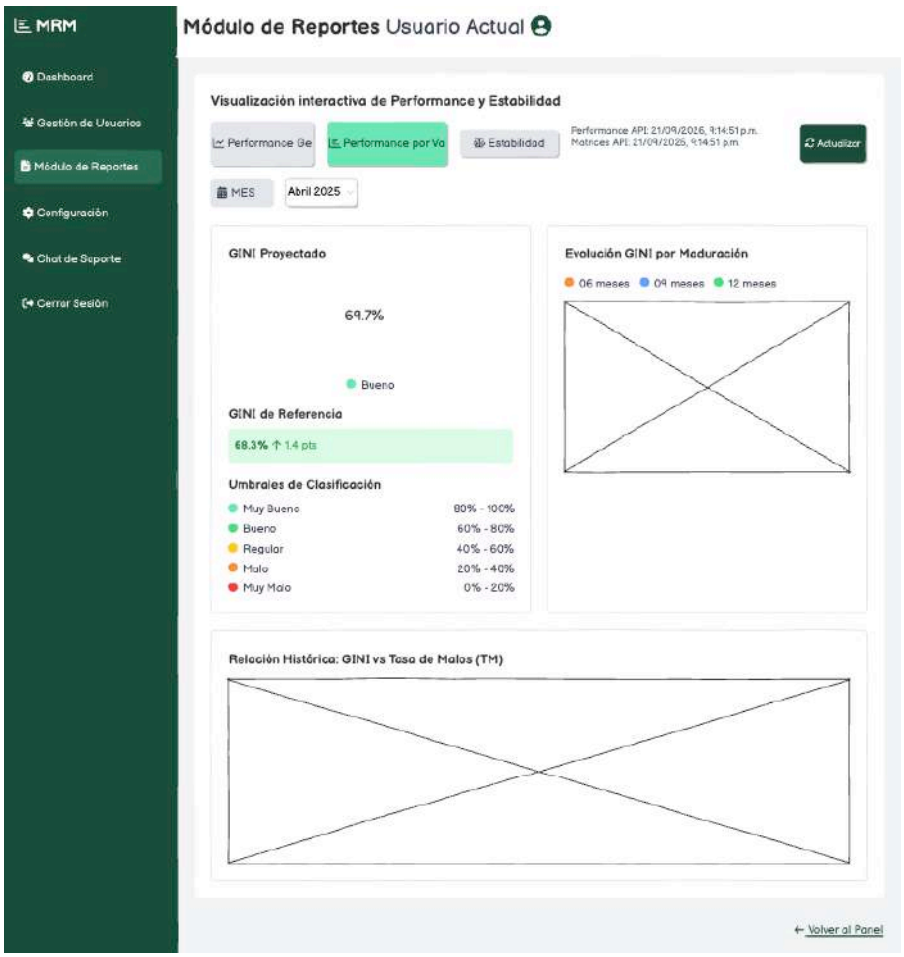


  - Aprobación de nuevos usuarios



- **Dashboards de Monitoreo**

  - Vista de Performance (Gini, KS, Tasas por Decil) y Vista de Estabilidad (PSI, Data Drift, Distribuciones)



- **Agente MRM (IA)**
    - Interfaz de Chat (Consultas de métricas y Normativa SBS)



3. Sistema de diseño

3.1 Paleta de colores

La interfaz combina una base formal (azul marino, blancos y grises fríos) con verde como color de confirmación. Los colores de riesgo se reservan para estados importantes y no deben utilizarse como decoración.

Token	HEX	Uso recomendado	Muestra
Texto principal / Navy	#111827	Títulos, valores KPI, texto de alta jerarquía	
Texto secundario	#475569	Descripciones, labels secundarios, fechas y notas	
Verde corporativo	#059669	Botones primarios, enlaces activos, indicadores positivos	
Verde oscuro	#047857	Hover/pressed, elementos de énfasis y encabezados semánticos	

Verde profundo	#064E3B	Fondos oscuros puntuales, como la pantalla de login	
Fondo general	#F8FAFC	Lienzo principal de dashboards y reportes	
Superficie / Card	#FFFFFF	Tarjetas, tablas, contenedores y formularios	
Borde neutro	#E2E8F0	Separadores, bordes de cards, inputs y tablas	
Estable	#10B981	Estado sin alerta / comportamiento estable	
Alerta temprana	#F59E0B	Riesgo intermedio, revisión preventiva	
Crítico	#EF4444	Alertas críticas, PSI fuera de umbral	

3.2 Semafoto de Riesgo

Los estados de riesgo para los modelos deben representarse con color + texto. Para PSI, las pantallas actuales emplean los siguientes colores:

Estado	Umbral PSI	Color	Tratamiento UI
Sin alerta	< 10%	Verde	Badge verde, icono de estado y mensaje positivo
Alerta media	10% - 25%	Ámbar	Badge ámbar y revisión preventiva
Alerta roja	> 25%	Rojo	Badge rojo, prioridad visual y revisión inmediata

3.3 Accesibilidad y diseño responsive

3.3.1 Accesibilidad

- Contraste mínimo recomendado de 4.5:1 para texto normal.
- No comunicar riesgo únicamente por color; incluir texto como “Sin Alerta”, “Alerta Ámbar” o “Alerta Roja”.
- Inputs con label asociado y mensajes de error descriptivos.
- Tablas con encabezados y orden de lectura coherente.
- Gráficos acompañados de valor textual y resumen para usuarios.

3.3.2 Breakpoints sugeridos

Breakpoint	Rango	Comportamiento
Desktop	≥ 1200 px	2-3 columnas, tablas completas, filtros en línea
Tablet	768-1199 px	Cards apilables, 2 columnas, tablas con scroll horizontal
Mobile	< 768 px	1 columna, filtros apilados, prioridad a KPIs y acciones esenciales

8.3 Reglas para tablas en móvil

Las tablas extensas no deben comprimirse hasta volver ilegible el contenido. Se recomienda scroll horizontal, columnas prioritarias y, cuando sea apropiado, una representación alternativa tipo card por registro.

ANEXO D. Documento de Pruebas de Software

Este anexo documenta los casos de prueba funcional de los principales casos de uso de la plataforma y la evaluación de atributos no funcionales mediante softwares especializados para cada tipo de prueba.

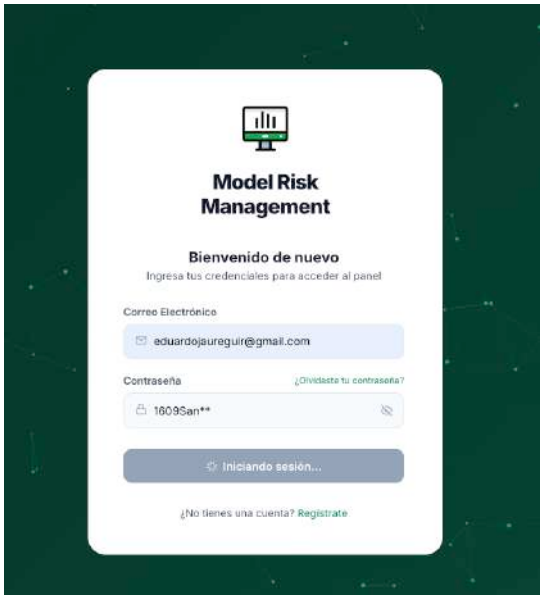
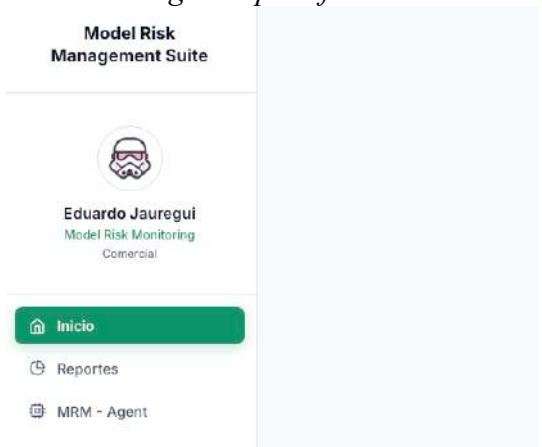
1. Alcance y estrategia

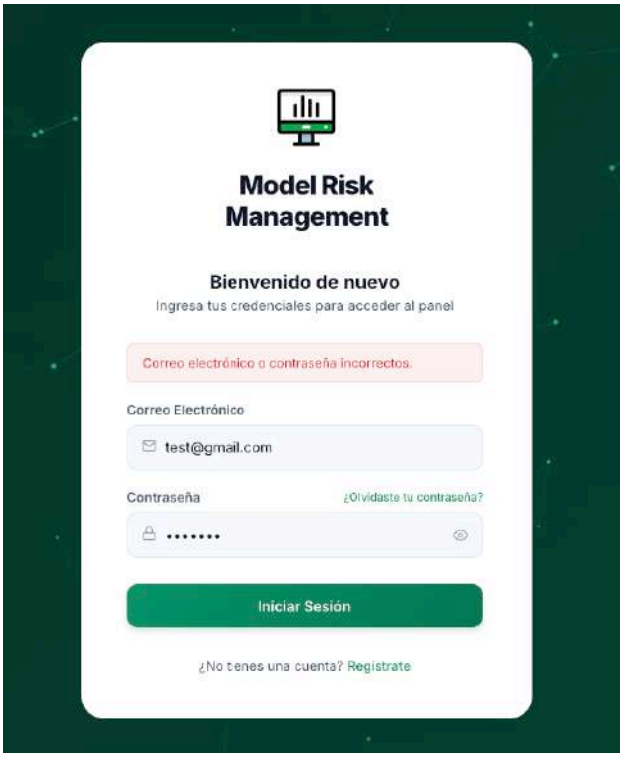
Las pruebas funcionales se documentan con una plantilla estándar por caso de uso a evaluar (casos de prueba funcional - CPF). Las pruebas no funcionales (PNF) emplean formatos propios, debido a que sus resultados se expresan mediante métricas, reportes y evidencias gráficas de software especializados en su medición.

2. Matriz de cobertura

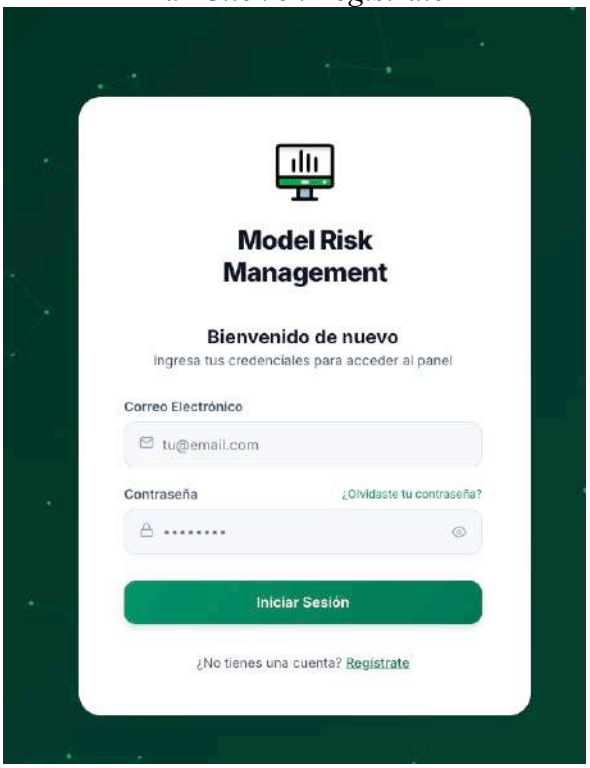
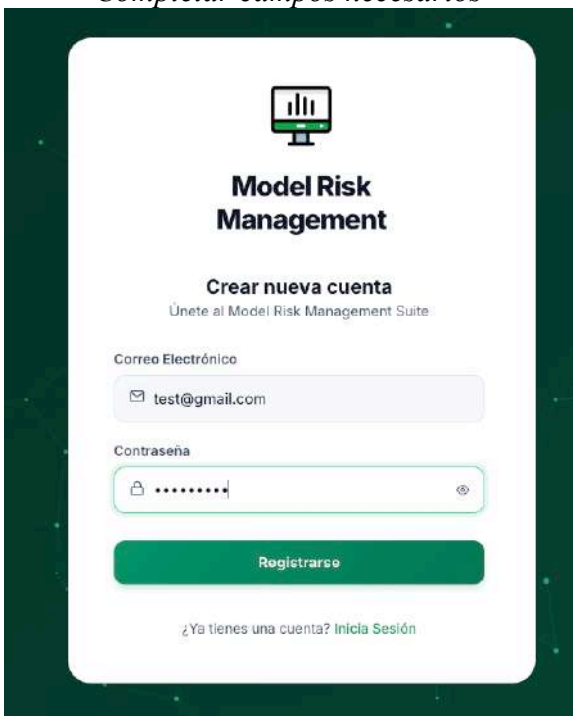
ID	Caso de uso / atributo	Prueba	Tipo
CPF-01	Inicio de Sesión	Autenticación y control de acceso	Funcional
CPF-02	Registro	Registro de nueva cuenta	Funcional
CPF-03	Configuración de usuario	Cambiar datos de usuario y actualizar	Funcional
CPF-04	Reportes de monitoreo	Interactuar con filtros del reporte	Funcional
CPF-05	Chat de Agente LLM	Iniciar nueva conversación Agente MRM	Funcional
PNF-01	Rendimiento web	Auditoría con PageSpeed Insights	No funcional
PNF-02	Pruebas de Carga y Stress	Prueba de carga y estrés concurrente	No funcional

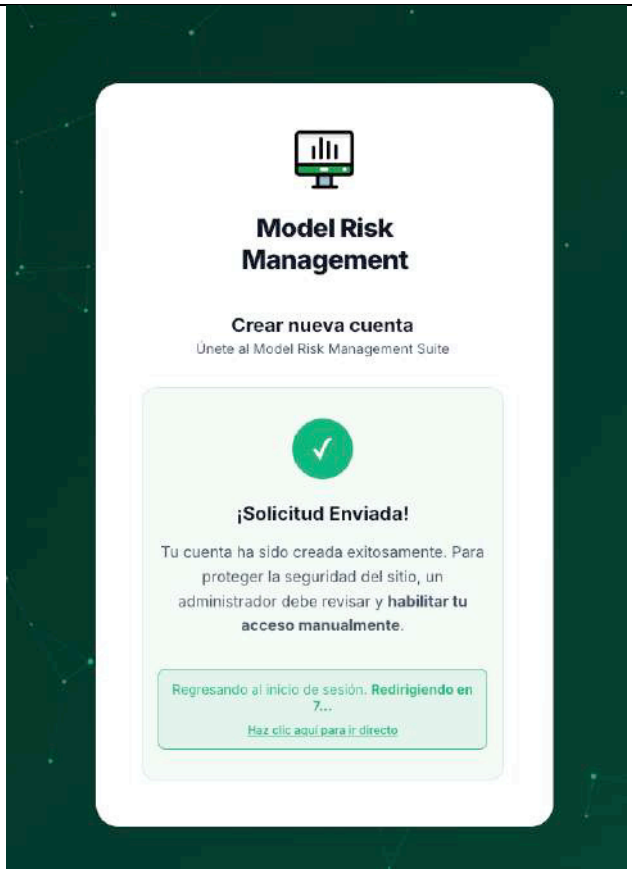
3. Casos de Prueba Funcionales

ID caso de prueba	CPF-01
Caso de uso relacionado	Autenticarse en la plataforma
Nombre	Autenticación y control de acceso
Descripción	Verificar el acceso de usuarios autorizados y la protección de rutas.
Precondiciones	Usuario habilitado en Cognito y plataforma disponible.
Pasos y condiciones de ejecución	• Abrir /login. • Ingresar credenciales válidas. • Acceder al panel. • Cerrar sesión. • Intentar volver a una ruta protegida. • Repetir con credenciales inválidas.
Resultado esperado	Se permite el acceso con credenciales válidas; al cerrar sesión o usar credenciales inválidas se impide el acceso.
Resultado obtenido	Exitoso
Evidencia	*Inicio de Sesión*  *Login en plataforma:*  *Cierre de sesión*

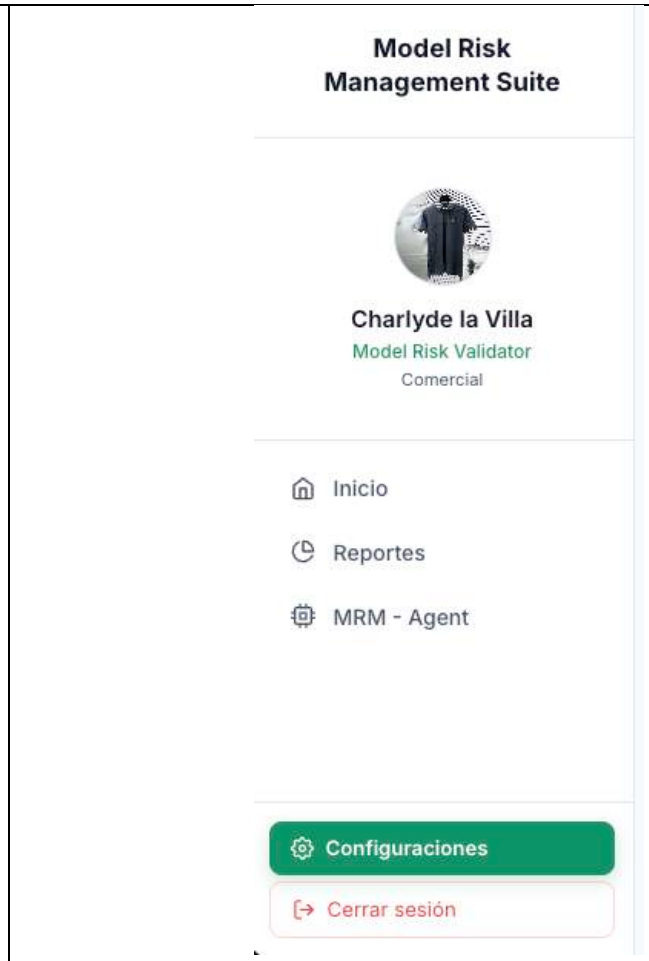
	 *Credenciales invalidas* 

ID caso de prueba	CPF-02
Caso de uso relacionado	Registrar Usuario
Nombre	Registro de nuevo usuario.
Descripción	Verificar que usuario creado se liste en pendientes.
Precondiciones	Plataforma disponible, usuario con cuenta no creada
Pasos y condiciones de ejecución	• Abrir página • Dar Click en botón regístrate • Completar los campos de correo y contraseña • Dar Click en botón de regístrate • Esperar pop-up de confirmación
Resultado esperado	La plataforma mandara mensaje que necesitamos aprobación de usuario

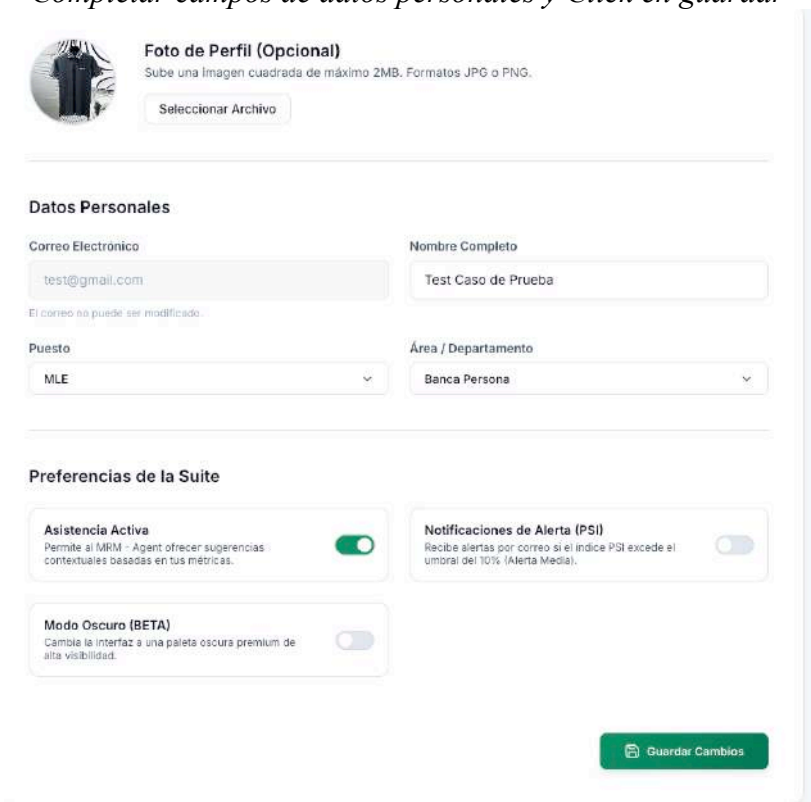
Resultado obtenido	Exitoso
Evidencia	*Dar Click en Regístrate*  *Completar campos necesarios*  *Pop-up de confirmación*

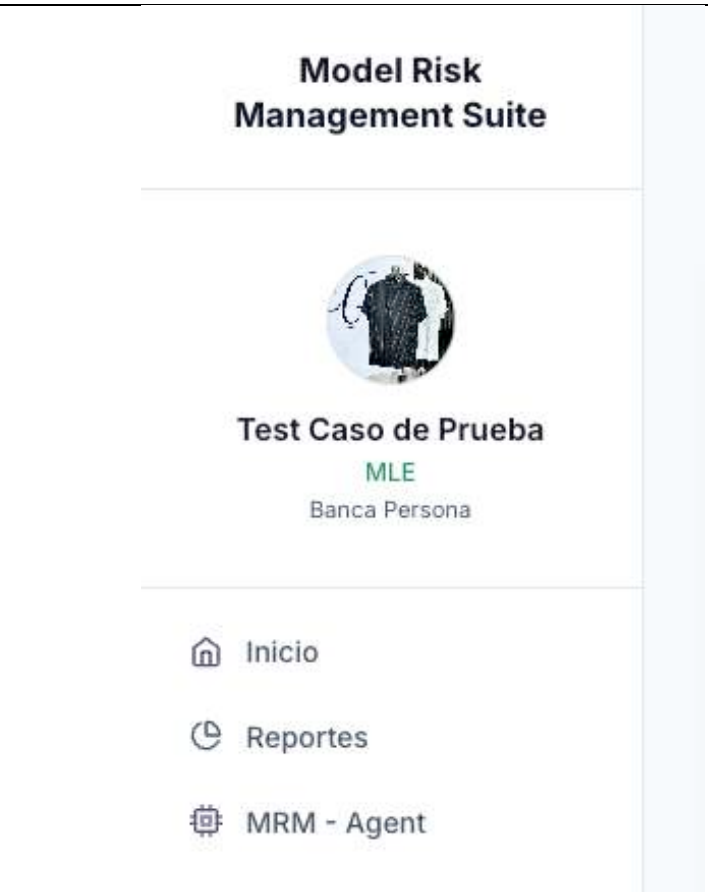
ID caso de prueba	CPF-03
Caso de uso relacionado	Configuración de Perfil de usuario
Nombre	Edición de datos de usuario
Descripción	Verificar que cambios en perfil se reflejen en la plataforma
Precondiciones	Usuario logueado y plataforma disponible
Pasos y condiciones de ejecución	• Dar Click en configuraciones • Editar Nombre Completo, puesto, área y cargar una foto de perfil • Dar Click en guardar • Refrescar la pagina
Resultado esperado	Los campos y la foto del usuario deberían actualizarse
Resultado obtenido	Exitoso
Evidencia	*Dar Click en configuraciones*



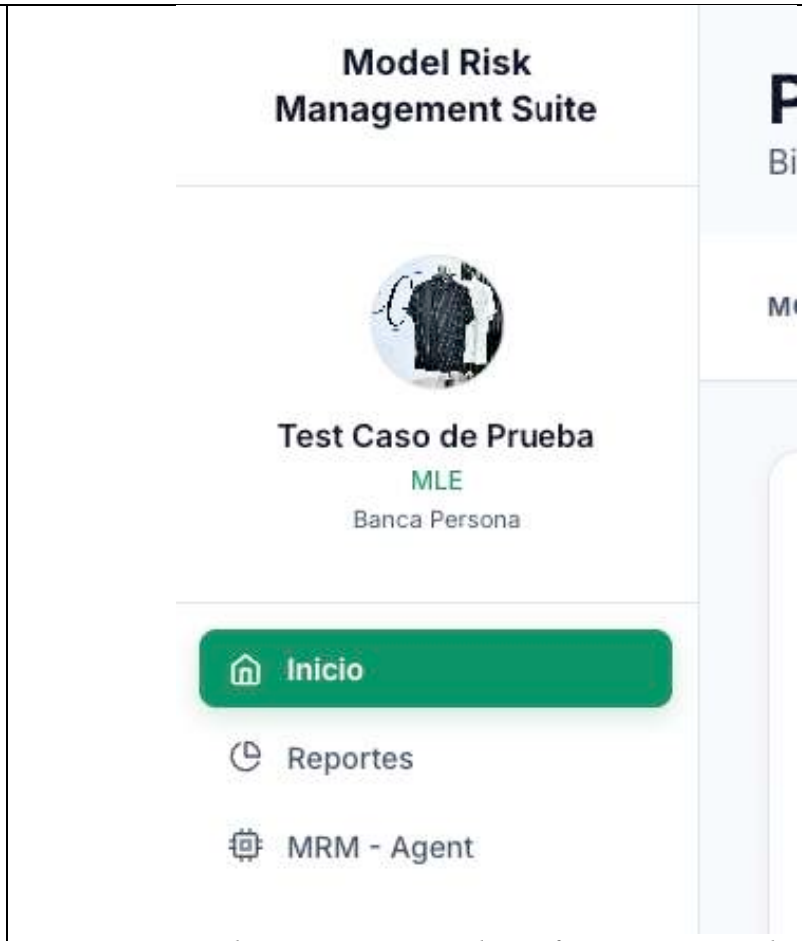
Completar campos de datos personales y Click en guardar



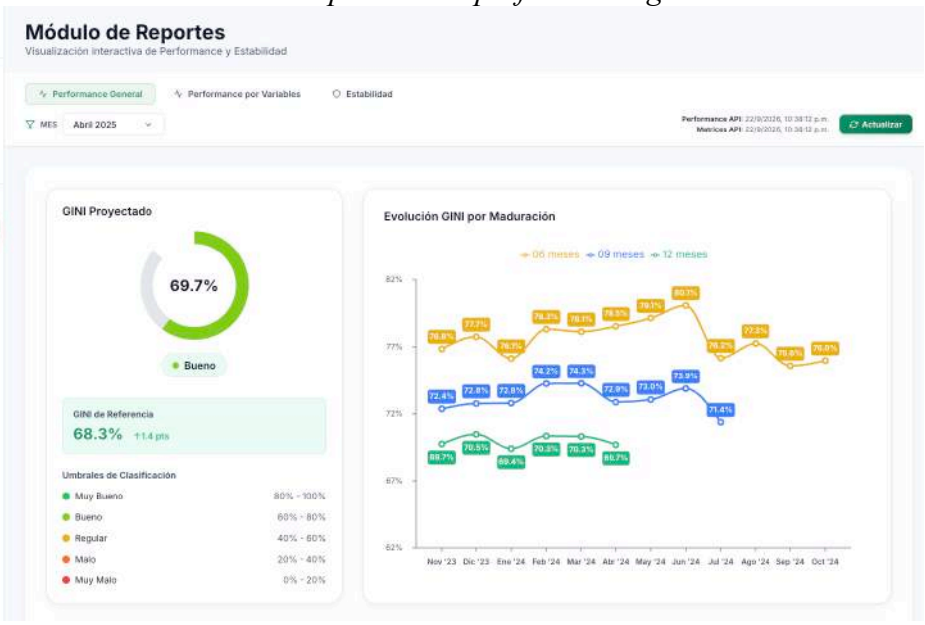
Refrescar la pagina

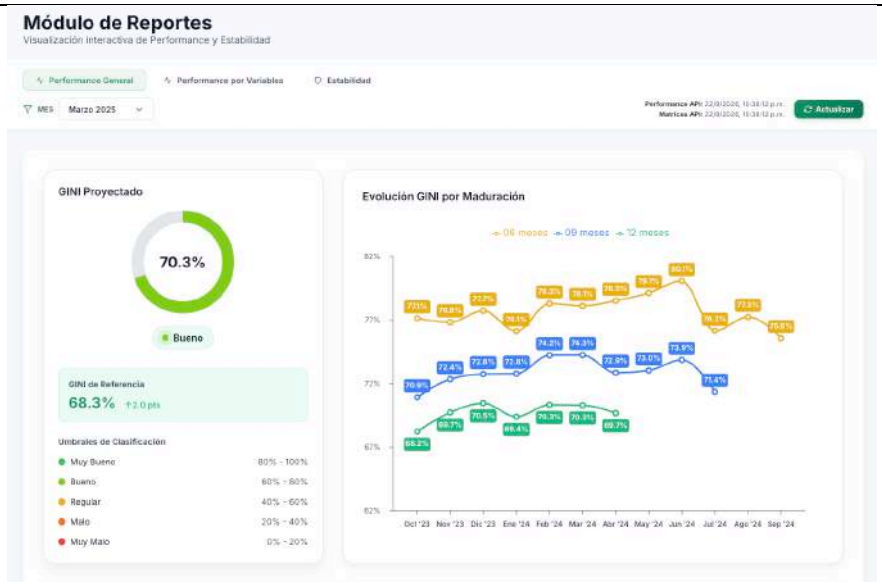
ID caso de prueba	CPF-04
Caso de uso relacionado	Edición de Filtros en Reportes
Nombre	Interacción el reporte y uso de filtros
Descripción	Se cambiará los filtros, pestañas de los reportes
Precondiciones	Usuario logueado y con permisos para ver reportes según su rol
Pasos y condiciones de ejecución	• Dar Click en Reportes. • Seleccionar la pestaña de performance general. • Cambiar el filtro a un mes anterior. • Seleccionar pestaña de performance por variables. • Cambiar el filtro a un mes anterior. • Seleccionar el filtro de estabilidad. • Cambiar el filtro a una variable especifica.
Resultado esperado	Las vistas deben actualizarse en función de los filtros seleccionados
Resultado obtenido	Exitoso
Evidencia	*Dar Click en Reportes*



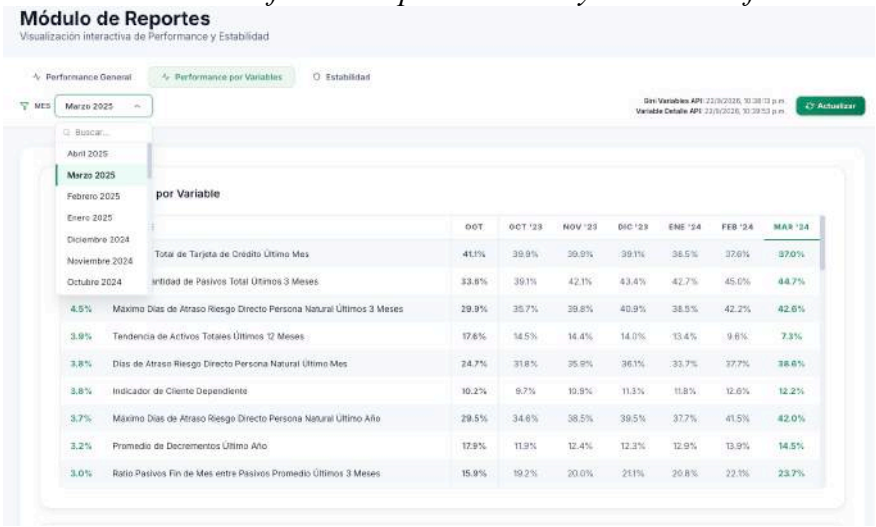
Seleccionar pestaña de performance general



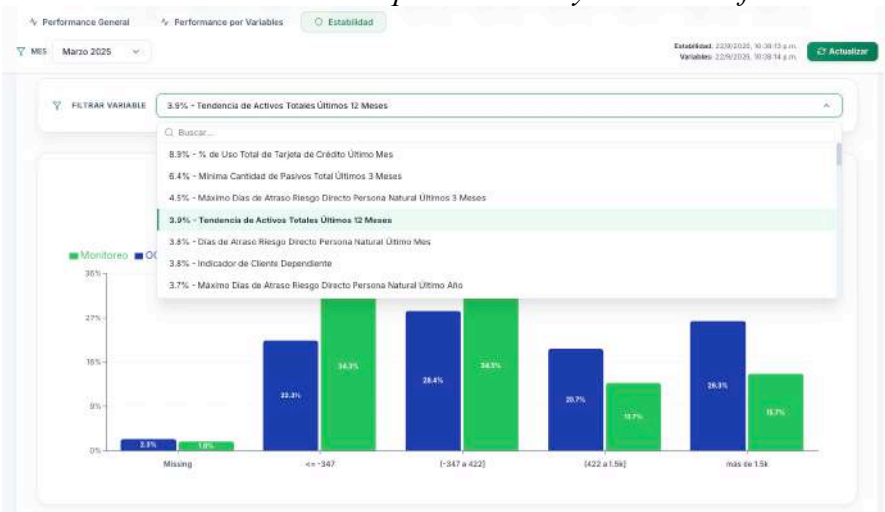
Cambiar filtro de mes a uno anterior

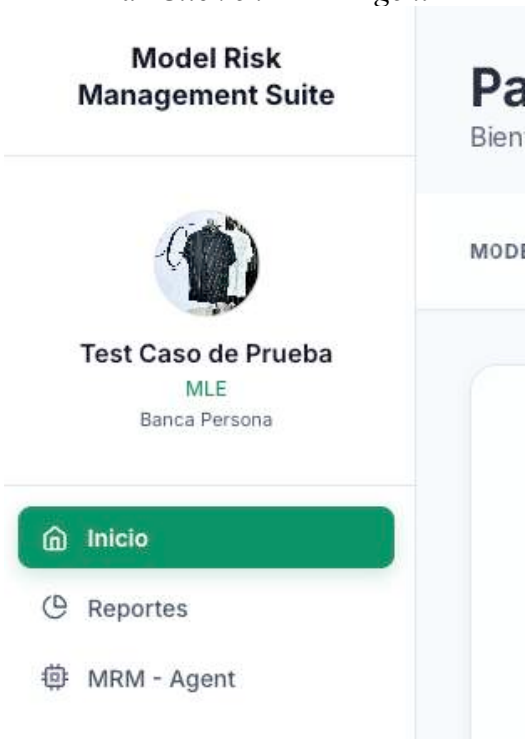


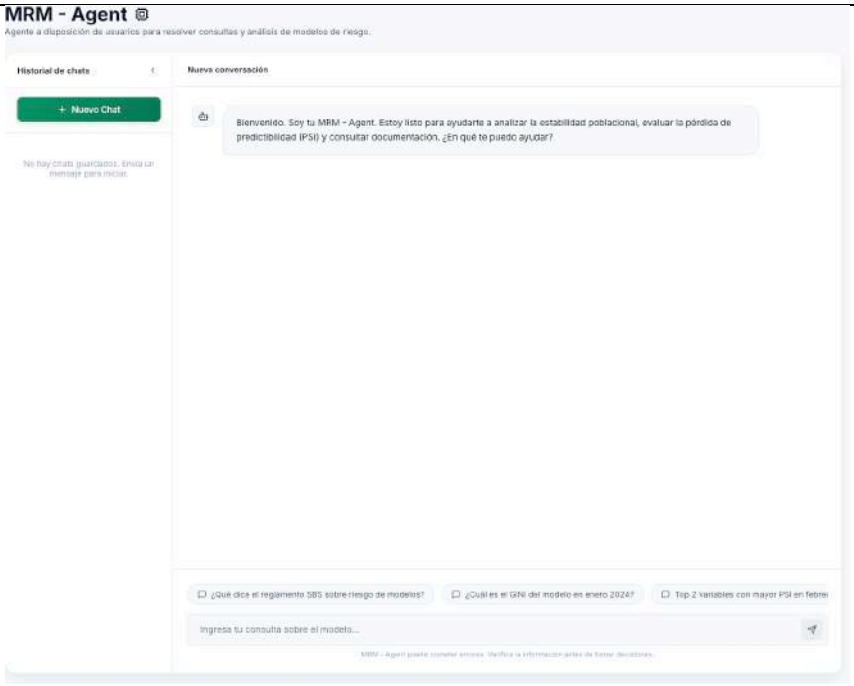
Seleccionar Performance por variables y cambiar el filtro



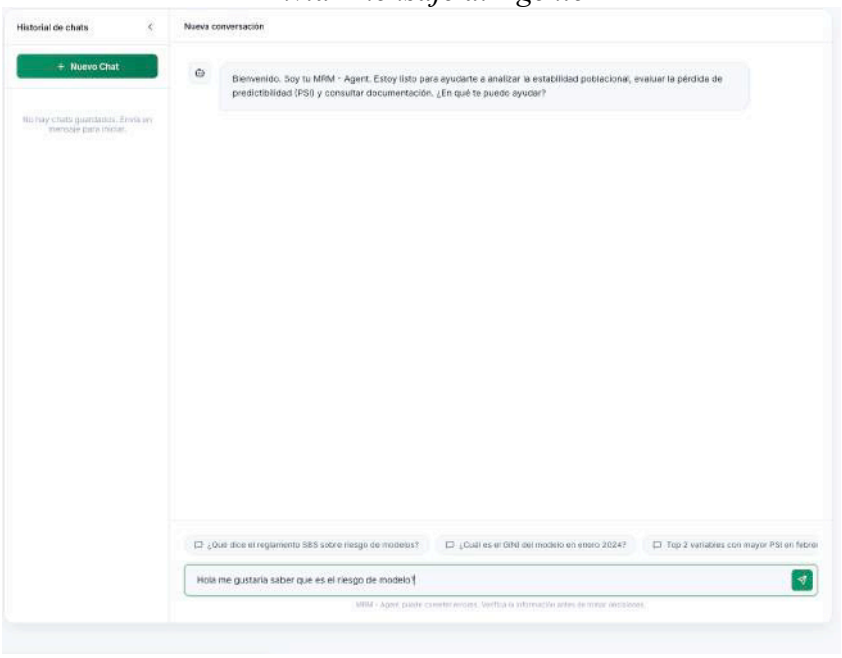
Seleccionar Estabilidad por variables y cambiar el filtro



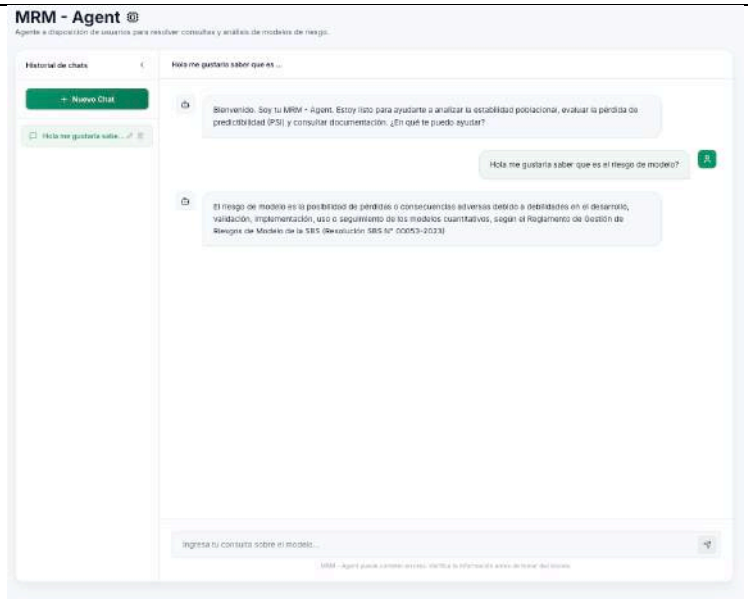
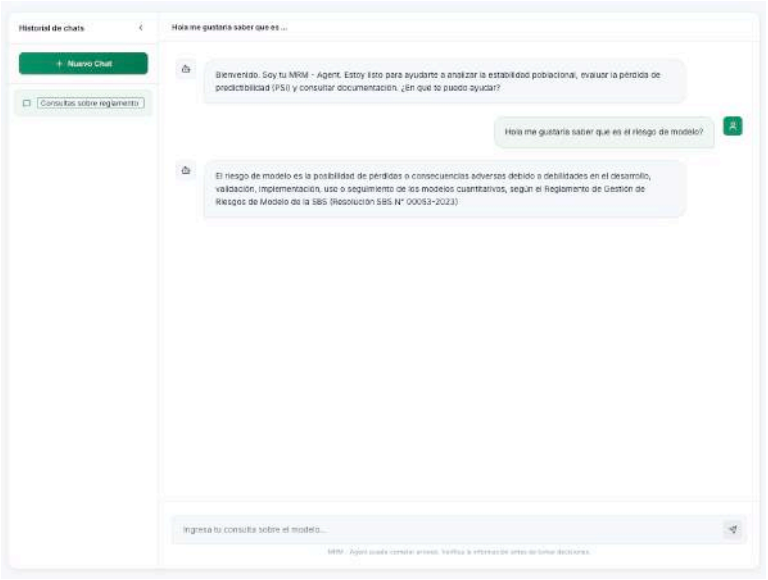
ID caso de prueba	CPF-05
Caso de uso relacionado	Creación de conversación con Agente LLM
Nombre	Uso del chat del Agente y Listado de Conversación/Edición
Descripción	Se mandará mensajes al agente esperando su respuesta
Precondiciones	Usuario logueado y con permisos para ver interactuar según su rol
Pasos y condiciones de ejecución	• Dar Click en MRM-Agent • Click en Nuevo Chat • Enviar un mensaje • Ir a menú desplegable de pestaña • Editar Titulo de la conversación
Resultado esperado	El chat debe crearse en la barra lateral y el agente debe responder las preguntas del usuario
Resultado obtenido	Exitoso
Evidencia	*Dar Click en MRM-Agent*  *Click en nuevo chat*



Enviar mensaje al Agente



Respuesta del Agente

	 *Edición de Conversación* 

4. Pruebas no funcionales

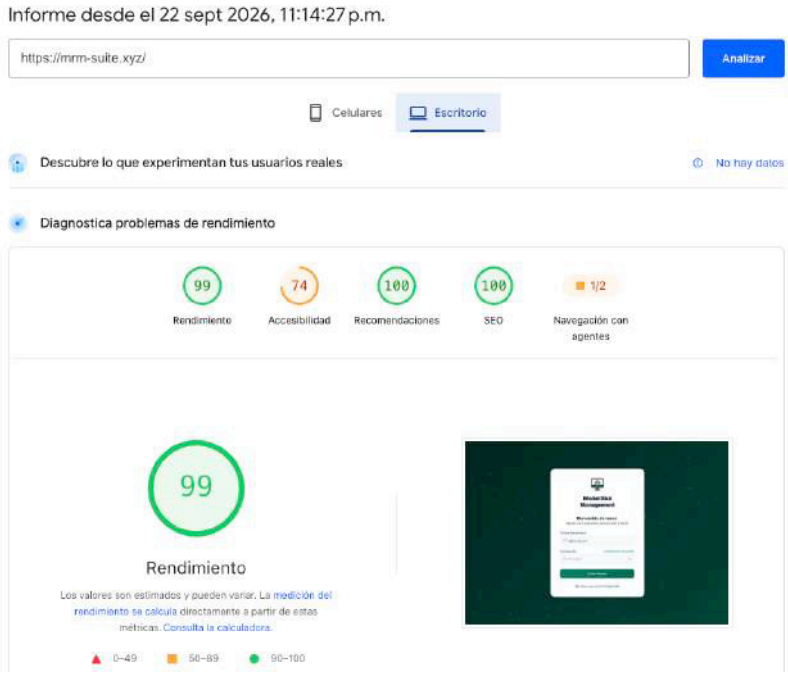
PNF-01. Auditoría de rendimiento y calidad web con PageSpeed Insights

- Objetivo: Evaluar calidad general de la página web (Rendimiento, Accesibilidad, Buenas prácticas, SEO y Web progresiva)
- Herramienta: PageSpeed Insights (Se tomará en cuenta los resultados de escritorio, ya que es el entorno donde más se utilizará el software)
- Criterio de aceptación mínimos:
    - Performance ≥ 80

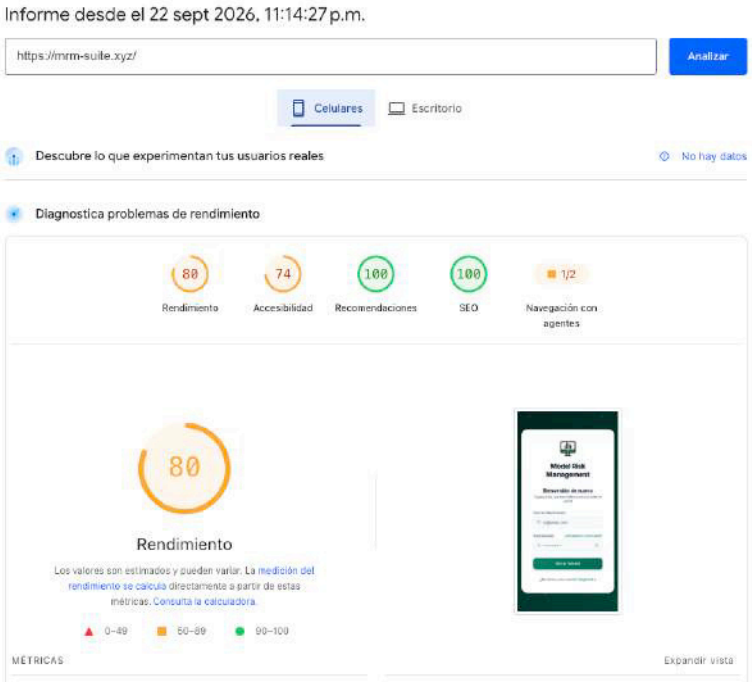
- Accessibility ≥ 70
- Best Practices ≥ 90
- SEO ≥ 90.

Performance	Accessibility	Best Practices	SEO
99/ 100	74/ 100	100/ 100	100/ 100

Resultados version Escritorio



Resultados versión Movil



PNF-02. Pruebas de carga y estrés concurrente de la plataforma web

Para las siguientes pruebas se uso el software de JMETER especializado en pruebas de carga y rendimiento.

Prueba de Carga

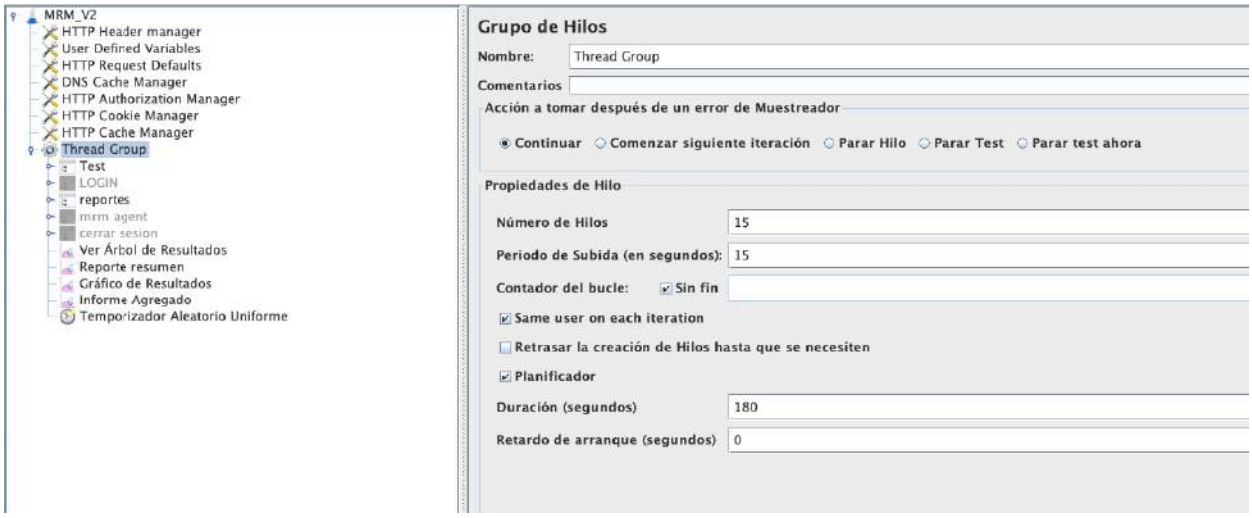
El objetivo es evaluar el comportamiento, la estabilidad operativa y los tiempos de respuesta del sistema bajo un escenario de concurrencia normal esperada, en el cual se interactúe bajo un flujo de un analista común. En la siguiente tabla se puede ver los supuestos considerados.

Parámetro	Valor Configurado	Justificación Técnica
Población Concurrente (Hilos)	15 usuarios virtuales	Representa la demanda concurrente estimada de analistas operando simultáneamente.
Tiempo de Rampa (Ramp-up)	15 segundos	Incorporación gradual de 1 usuario por segundo.
Duración de la Prueba	180 segundos (3 minutos)	Ventana temporal suficiente para sostener a los 15 usuarios en régimen constante.
Modo de Ejecución	Bucle infinito + Planificador activo	Mantiene la concurrencia activa y fija el cese a los 180 segundos.
Tiempo de Reflexión (Think Time)	Uniform Random Timer: 5,000 ms a 10,000 ms	Modela el comportamiento analítico humano (lectura de métricas/gráficos antes de hacer la siguiente acción) y evita saturación artificial.

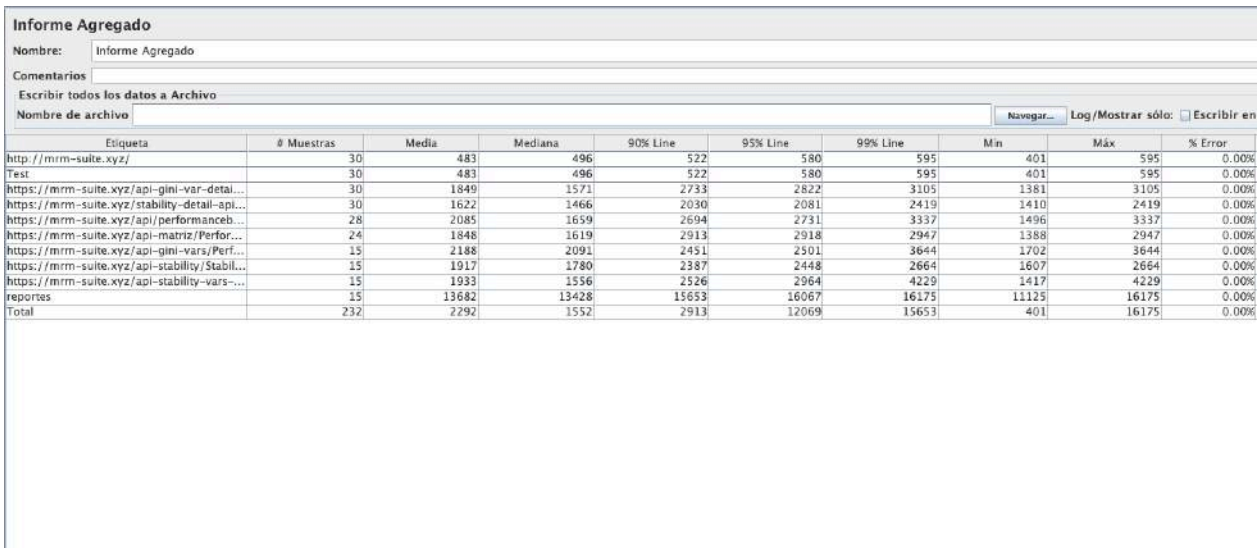
En la prueba se ejecutaron 232 transacciones sin un solo fallo HTTP (tasa de error del 0.00%). El sistema demostró capacidad completa de absorción de la carga normal proyectada (15 usuarios en simultáneo). Las APIs individuales de cómputo analítico (Gini, Estabilidad, Matriz de Confusión y PSI) mostraron una media homogénea de entre 1,622 ms y 2,188 ms. La mediana global se situó en 1,552 ms y el Percentil 90 en 2,913 ms, lo que demuestra que el 90% de las solicitudes se resolvieron en menos de 3 segundos, manteniéndose en rangos aceptables para procesos intensivos de backend.

El entorno analizado aprueba satisfactoriamente la prueba de carga bajo condiciones normales, evidenciando robustez, ausencia de fugas inmediatas y suficiente capacidad de concurrencia para el flujo operativo diario. En las siguientes figuras se puede apreciar la evidencia de las pruebas realizadas

Configuración de Hilos



Informe Agregado



Etiqueta	# Muestras	Media	Mediana	90% Line	95% Line	99% Line	Min	Máx	% Error
http://mrm-suite.xyz/	30	483	496	522	580	595	401	595	0.00%
Test	30	483	496	522	580	595	401	595	0.00%
https://mrm-suite.xyz/api-gini-var-detai...	30	1849	1571	2733	2822	3105	1381	3105	0.00%
https://mrm-suite.xyz/stability-detail-api...	30	1622	1466	2030	2081	2419	1410	2419	0.00%
https://mrm-suite.xyz/api/performanceb...	28	2085	1659	2694	2731	3337	1496	3337	0.00%
https://mrm-suite.xyz/api-matriz/Perfor...	24	1848	1619	2913	2918	2947	1388	2947	0.00%
https://mrm-suite.xyz/api-gini-vars/Perf...	15	2188	2091	2451	2501	3644	1702	3644	0.00%
https://mrm-suite.xyz/api-stability/Stabil...	15	1917	1780	2387	2448	2664	1607	2664	0.00%
https://mrm-suite.xyz/api-stability-vars-...	15	1933	1556	2526	2964	4229	1417	4229	0.00%
reportes	15	13682	13428	15653	16067	16175	11125	16175	0.00%
Total	232	2292	1552	2913	12069	15653	401	16175	0.00%

Gráfico de Resultados



Prueba de Estrés

El objetivo es evaluar la estabilidad, los tiempos de respuesta ante un pico de tráfico extraordinario del 100% sobre la demanda base (duplicando de 15 a 30 usuarios concurrentes), con el propósito de determinar la capacidad máxima tolerable. En la siguiente tabla se puede ver los supuestos considerados.

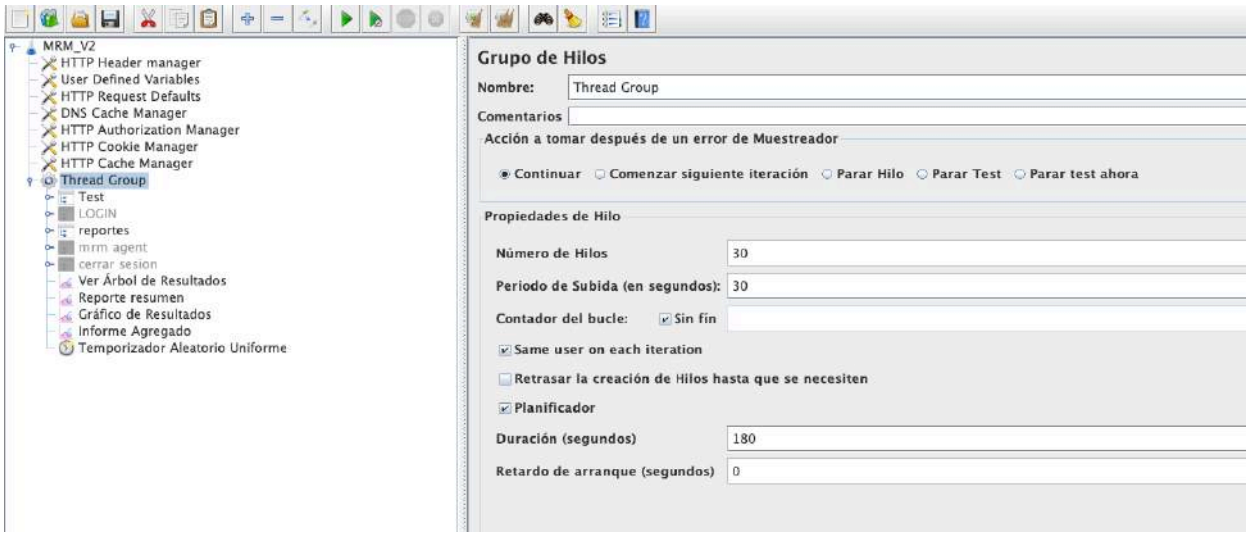
Parámetro	Valor Configurado	Justificación Técnica
Población Concurrente (Hilos)	30 usuarios virtuales	Simula un evento excepcional de alta concurrencia en simultáneo (el doble del tráfico diario habitual)
Tiempo de Rampa (Ramp-up)	30 segundos	Incorporación gradual de 1 usuario por segundo.
Duración de la Prueba	180 segundos (3 minutos)	Ventana temporal suficiente para sostener a los 15 usuarios en régimen constante.
Modo de Ejecución	Bucle infinito + Planificador activo	Mantiene la concurrencia activa y fija el cese a los 180 segundos.

Tiempo de Reflexión (Think Time)	Uniform Random Timer: 3,000 ms a 8,000 ms	Modela el comportamiento analítico humano (lectura de métricas/gráficos antes de hacer la siguiente acción) y evita saturación artificial.

La mayoría de los microservicios analíticos (Gini Var Detail, Stability Detail, Matriz de Confusión y PSI) operaron con 0.00% de error y latencias medias contenidas entre 1.6 y 1.7 segundos. También se observó que sube durante los primeros 30 segundos (fase de entrada) y luego se estabiliza de forma horizontal alrededor de los 2,391 ms. El servidor no sufrió asfixia descontrolada.

Con 30 usuarios, reportes promedió 12,796 ms (12.7 s) y tuvo un 5.00% de error (solo 3 fallos en 60 intentos). La infraestructura actual garantiza una operación impecable para hasta 15-20 analistas concurrentes, tolera ráfagas pico de hasta 30 analistas con degradación mínima (1.1% error), y requiere escalado de recursos si se prevé mucho más allá de los 30 usuarios en simultáneo. Bajo la casuística actual en un equipo de 10 personas como máximo, la plataforma es más que apta para su funcionamiento. En las siguientes figuras se puede apreciar la evidencia de las pruebas realizadas

Configuración de Hilos



Informe Agregado

Informe Agregado

Nombre: Informe Agregado

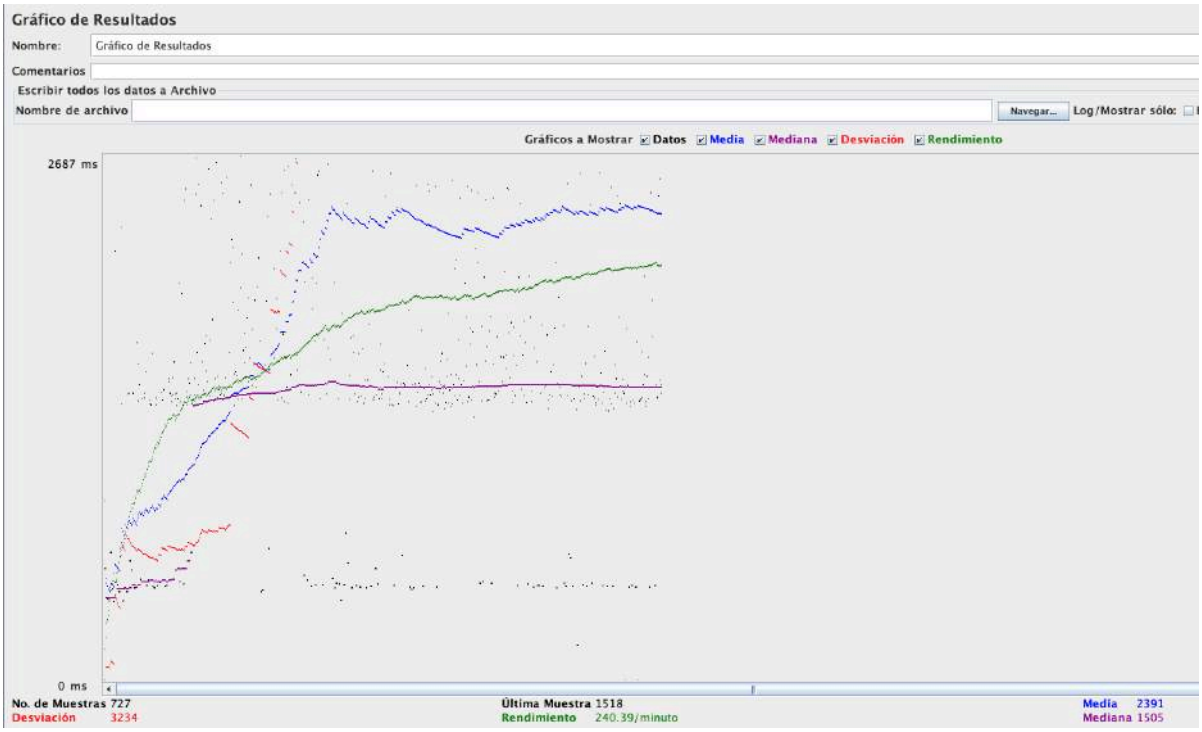
Comentarios

Escribir todos los datos a Archivo

Nombre de archivo: Navegar... Log/Mostrar sólo: Escribir en

Etiqueta	# Muestras	Media	Mediana	90% Line	95% Line	99% Line	Mín	Máx	% Error
http://mrm-suite.xyz/	88	508	488	578	651	759	182	1475	1.14%
Test	90	497	488	578	642	759	0	1475	1.11%
https://mrm-suite.xyz/api-gini-var-detai...	85	1607	1470	2200	2509	2765	1377	3088	0.00%
https://mrm-suite.xyz/stability-detail-api...	80	1624	1447	2438	2460	2818	1386	3033	0.00%
https://mrm-suite.xyz/api/performanceb...	74	1925	1569	2687	2731	3218	162	3533	1.35%
https://mrm-suite.xyz/api-matriz/Perfor...	68	1698	1475	2502	2898	2929	1352	3111	0.00%
https://mrm-suite.xyz/api-gini-vars/Perf...	62	2051	1946	2758	2823	3023	174	3623	1.61%
https://mrm-suite.xyz/api-stability/Stabil...	60	2030	1692	2846	2964	4707	172	5231	1.67%
https://mrm-suite.xyz/api-stability-vars-...	60	1788	1519	2603	2673	2860	1423	3716	0.00%
reportes	60	12796	12491	14898	15815	16728	9643	17470	5.00%
Total	727	2391	1505	2950	12079	14832	0	17470	1.10%

Gráfico de Resultados



ANEXO E. Validación del Artefacto Modelo probabilidad de impago

VALIDACIÓN DEL ARTEFACTO

Modelo de Probabilidad de Impago

1. Introducción

La validación del artefacto de modelo es un proceso sistemático que verifica si un modelo de riesgo crediticio cumple con los requisitos establecidos y satisface las necesidades de los usuarios finales. Este documento presenta la validación del Modelo Comportamental PP de Préstamos Personales.

El modelo fue desarrollado empleando técnicas de ML y aprendizaje automático supervisado, integrando variables comportamentales del portafolio de préstamos personales. La validación fue realizada por un consultor externo independiente al equipo de desarrollo, evaluando el modelo en seis frentes principales de análisis cuantitativo y cualitativo.

El propósito de esta validación es evaluar la calidad del modelo desde diversas perspectivas, incluyendo la suficiencia de la documentación, la calidad de los datos, la capacidad de réplica, la metodología estadística empleada, la representatividad de las variables de desarrollo y la bondad de ajuste del modelo.

2. Descripción del Artefacto

El Modelo Comportamental PP de Préstamos Personales es un sistema de scoring crediticio diseñado para estimar la probabilidad de incumplimiento (PD) de clientes con préstamos personales vigentes. El modelo integra información del comportamiento de pago histórico, características demográficas y variables macroeconómicas relevantes.

2.1 Componentes Principales del Modelo

El artefacto se compone de los siguientes módulos técnicos:

- Módulo de Preprocesamiento de Datos: Limpieza, transformación y validación de las variables de entrada del modelo.
- Módulo de Estimación: Algoritmo de ML con selección de variables mediante análisis de Information Value (IV) y Weight of Evidence (WoE).

- Módulo de Validación Estadística: Cálculo de métricas de desempeño (Gini, KS, AUC-ROC).
- Módulo de Monitoreo: Dashboard de seguimiento continuo del comportamiento del modelo en producción.

2.2 Alcance de la Validación

La validación cubre el ciclo completo de desarrollo del modelo, desde la definición del universo de análisis hasta la evaluación del desempeño predictivo. Se excluyen del alcance los procesos de integración tecnológica con los sistemas core de la institución.

3. Proceso de Validación

El proceso de validación fue conducido mediante una revisión exhaustiva e independiente de toda la documentación técnica y los entregables del modelo. La metodología empleada se basó en los estándares de validación de modelos de riesgo crediticio, siguiendo las mejores prácticas regulatorias del sector financiero.

A continuación, se presentan los criterios de evaluación empleados por el consultor externo para determinar la conformidad del modelo en cada uno de los frentes de validación:

Criterio de Evaluación	Valoración
Completitud de la documentación del modelo	[ ] Cumple [ ] No cumple
Calidad y representatividad de los datos	[ ] Cumple [ ] No cumple
Correcta replicación del modelo	[ ] Cumple [ ] No cumple
Adecuación de la metodología estadística	[ ] Cumple [ ] No cumple
Representatividad de las variables de desarrollo	[ ] Cumple [ ] No cumple
Bondad de ajuste del modelo	[ ] Cumple [ ] No cumple

4. Resultados de la Validación

Tras la revisión exhaustiva del modelo, el consultor externo emitió los resultados de validación para cada uno de los seis ámbitos evaluados. Los resultados se presentan en la siguiente tabla:

Ámbito de Evaluación	Resultado de Validación
Documentación	✓ Sin alerta
Data Quality	✓ Sin alerta
Réplica	✓ Sin alerta
Metodología empleada	✓ Sin alerta
Variables muestra de desarrollo	✓ Sin alerta
Bondad de Ajuste	✓ Sin alerta

4.1 Resumen de Puntos Clave y Plan de Acción

Punto observado: Se recomienda detallar cada una de las variables evaluadas y el criterio de descarte para un mejor seguimiento.

Riesgo: Bajo.

Punto fijo: El equipo de modelamiento redefinirá el mínimo de variables para las cuales se enviará el detalle, según su tier.

Punto observado: Los consultores externos no pudieron recrear completamente las bases de datos por falta de accesos. Se revisaron detalladamente las lógicas y bases compartidas. El equipo interno de validación complementa estas validaciones con sus procesos actuales. Se evalúa el acceso futuro a los esquemas de desarrollo para futuras validaciones.

Riesgo: Bajo.

✓ Sin alerta — Riesgo bajo en todos los ámbitos evaluados

5. Conclusión y Conformidad

En base a los resultados obtenidos en el proceso de validación, el consultor externo concluye que el Modelo Comportamental PP de Préstamos Personales se encuentra apto para su puesta en producción, habiendo superado satisfactoriamente todos los frentes de validación requeridos.

El modelo cumple con los estándares metodológicos, la calidad de datos es adecuada, la réplica fue exitosa dentro de los márgenes establecidos, y las métricas de bondad de ajuste se encuentran dentro de los rangos aceptables para este tipo de modelo.

Las recomendaciones emitidas son de carácter informativo y no representan impedimento para la implementación del modelo. Se sugiere incorporarlas en el plan de mejora continua y abordarlas en los próximos ciclos de revisión del modelo.

Conclusión de la Validación: Se concluye que el Modelo Comportamental PP de Préstamos Personales está apto para su puesta en producción, habiendo superado satisfactoriamente todos los frentes de validación establecidos por el equipo consultor.

ANEXO F. Validación de la Plataforma Web MRM Suite (Experto 01)

VALIDACIÓN DEL ARTEFACTO

Plataforma de Monitoreo de Modelos de Machine Learning

Información del Evaluador	
Nombres y apellidos:	Luis Asmat Cueva
Título / Profesión:	Ingeniero de Computación y Sistemas
Grado académico:	Ingeniero
Años de experiencia en el campo relacionado con la profesión:	14 años
Área(s) de especialización:	Machine Learning
Cargo en la empresa/organización:	Lead MRM

1. Introducción

La validación del artefacto de software es un proceso sistemático que verifica si un sistema cumple con los requisitos establecidos y satisface las necesidades de sus usuarios. Este documento presenta la validación de la Plataforma de Monitoreo de Modelos de Machine Learning, desarrollada para el equipo de Model Risk Management.

La plataforma permite visualizar métricas de performance y estabilidad de modelos, analizar clientes, interpretar variables mediante gráficos SHAP, y consultar información a través de un agente LLM integrado. El acceso está restringido por roles y la plataforma se encuentra desplegada en AWS, con ingreso únicamente mediante usuario creado.

El propósito de esta validación es evaluar la calidad de la plataforma desde la perspectiva de los usuarios finales del equipo de Model Risk Management, considerando los criterios de usabilidad, rendimiento, funcionalidad, seguridad y satisfacción general, en concordancia con el estándar ISO/IEC 25010.

2. Descripción del Artefacto

La Plataforma de Monitoreo de Modelos de ML es una aplicación web desplegada en AWS que centraliza el seguimiento y análisis de los modelos de riesgo crediticio en producción. Sus principales módulos son:

- Métricas de Performance: Visualización de indicadores clave como Gini, KS y AUC-ROC por modelo.
- Estabilidad de Modelos: Análisis de PSI (Population Stability Index) y CSI (Characteristic Stability Index) para detectar degradación.
- Análisis de Clientes: Consulta y filtrado de información de clientes vinculados a los modelos.
- Interpretabilidad SHAP: Gráficos de contribución de variables para la explicación de predicciones individuales y globales.
- Agente LLM: Asistente conversacional para consultas sobre métricas, documentación y comportamiento de los modelos.
- Control de Acceso por Roles: Gestión de permisos que restringe el acceso a módulos según el perfil del usuario.

3. Escala de Evaluación

Para la evaluación de la plataforma se utilizó la escala Likert de 5 puntos que se presenta en la siguiente tabla:

Puntaje	Significado
1	Muy Deficiente
2	Deficiente
3	Aceptable
4	Bueno
5	Muy Bueno

Los criterios y preguntas de validación empleados para la evaluación se presentan en la siguiente tabla:

Aspecto	Pregunta de Evaluación
Usabilidad	La plataforma es intuitiva y fácil de navegar para usuarios con distintos roles.
Rendimiento	La plataforma responde con rapidez al consultar métricas, gráficos y análisis de modelos.
Funcionalidad	Las herramientas de monitoreo, el agente LLM y el análisis SHAP funcionan correctamente.
Seguridad	Confío en que el control de acceso por roles protege adecuadamente la información.
Satisfacción	Estoy satisfecho con la plataforma como solución para el monitoreo de modelos de ML.

4. Tabla de Validación de Criterios

La norma ISO/IEC 25010 establece criterios de calidad de software tales como usabilidad, funcionalidad, rendimiento y seguridad para evaluar la calidad de un producto de software. A continuación, se presenta la tabla de validación que incluye preguntas específicas para cada criterio evaluado en la plataforma de monitoreo:

Criterio / Pregunta de Validación	Evaluación (1 a 5)
Usabilidad	
¿La plataforma es intuitiva y fácil de usar para usuarios con distintos roles (analistas, supervisores)?	5
¿La navegación entre los diferentes módulos (métricas, estabilidad, clientes, SHAP) es clara y fluida?	4
¿El diseño visual facilita la lectura e interpretación de los gráficos y resultados?	4
¿El sistema de roles y permisos es fácil de entender para el usuario final?	4
¿La plataforma se adapta correctamente a la resolución y dispositivo del usuario?	4

Rendimiento	
¿La plataforma responde con rapidez al consultar métricas de performance de los modelos?	5
¿Las visualizaciones de estabilidad (PSI, CSI) cargan sin demoras significativas?	4
¿El agente LLM responde en tiempos aceptables ante consultas sobre los modelos?	4
¿El acceso al análisis SHAP e interpretación de variables es ágil?	3
Funcionalidad	
¿Las métricas de performance (Gini, KS, AUC) se muestran correcta y completamente?	5
¿El módulo de análisis de clientes permite filtrar y consultar información de manera efectiva?	4
¿El agente LLM responde de forma precisa y coherente a preguntas sobre los modelos?	4
¿Los gráficos de SHAP representan adecuadamente la contribución de cada variable?	3
¿El control de acceso por roles restringe correctamente los módulos según el perfil del usuario?	5
Seguridad	
¿El acceso a la plataforma mediante usuario y contraseña (desplegada en AWS) es seguro y confiable?	5
¿Los usuarios solo pueden ver la información correspondiente a su rol asignado?	5
¿Confío en que la información sensible de los modelos y clientes está protegida en la plataforma?	5
Satisfacción General	
¿La plataforma cubre las necesidades de monitoreo de modelos de ML del equipo de Model Risk Management?	4
¿Recomendaría el uso de esta plataforma como herramienta de seguimiento de modelos?	5
¿Considera que la plataforma es un aporte valioso para el proceso de validación y monitoreo de modelos?	4
Evalúe de manera general la plataforma de monitoreo de modelos de ML.	4

5. Comentarios del Evaluador (Opcional)

-

6. Confirmación de Aprobación

El presente documento ha sido revisado y completado por el evaluador, quien certifica la validación de la Plataforma de Monitoreo de Modelos de Machine Learning.

Fecha:	02 de mayo de 2026
Firma:	

ANEXO G. Validación de la Plataforma Web MRM Suite (Experto 02)

VALIDACIÓN DEL ARTEFACTO

Plataforma de Monitoreo de Modelos de Machine Learning

Información del Evaluador	
Nombres y apellidos:	Jesus Zelaya Contreras
Título / Profesión:	Licenciado en Matemáticas
Grado académico:	Licenciado
Años de experiencia en el campo rclacionado con la profesión:	4 años
Área(s) de especialización:	Inteligencia Artificial
Cargo en la empresa/organización:	Data Scentists

1. Introducción

La validación del artefacto de software es un proceso sistemático que verifica si un sistema cumple con los requisitos establecidos y satisface las necesidades de sus usuarios. Este documento presenta la validación de la Plataforma de Monitoreo de Modelos de Machine Learning, desarrollada para el equipo de Model Risk Management.

La plataforma permite visualizar métricas de performance y estabilidad de modelos, analizar clientes, interpretar variables mediante gráficos SHAP, y consultar información a través de un agente LLM integrado. El acceso está restringido por roles y la plataforma se encuentra desplegada en AWS, con ingreso únicamente mediante usuario creado.

El propósito de esta validación es evaluar la calidad de la plataforma desde la perspectiva de los usuarios finales del equipo de Model Risk Management, considerando los criterios de usabilidad, rendimiento, funcionalidad, seguridad y satisfacción general, en concordancia con el estándar ISO/IEC 25010.

2. Descripción del Artefacto

La Plataforma de Monitoreo de Modelos de ML es una aplicación web desplegada en AWS que centraliza el seguimiento y análisis de los modelos de riesgo crediticio en producción. Sus principales módulos son:

- Métricas de Performance: Visualización de indicadores clave como Gini, KS y AUC-ROC por modelo.
- Estabilidad de Modelos: Análisis de PSI (Population Stability Index) y CSI (Characteristic Stability Index) para detectar degradación.
- Análisis de Clientes: Consulta y filtrado de información de clientes vinculados a los modelos.
- Interpretabilidad SHAP: Gráficos de contribución de variables para la explicación de predicciones individuales y globales.
- Agente LLM: Asistente conversacional para consultas sobre métricas, documentación y comportamiento de los modelos.
- Control de Acceso por Roles: Gestión de permisos que restringe el acceso a módulos según el perfil del usuario.

3. Escala de Evaluación

Para la evaluación de la plataforma se utilizó la escala Likert de 5 puntos que se presenta en la siguiente tabla:

Puntaje	Significado
1	Muy Deficiente
2	Deficiente
3	Aceptable
4	Bueno
5	Muy Bueno

Los criterios y preguntas de validación empleados para la evaluación se presentan en la siguiente tabla:

Aspecto	Pregunta de Evaluación
Usabilidad	La plataforma es intuitiva y fácil de navegar para usuarios con distintos roles.
Rendimiento	La plataforma responde con rapidez al consultar métricas, gráficos y análisis de modelos.
Funcionalidad	Las herramientas de monitoreo, el agente LLM y el análisis SHAP funcionan correctamente.
Seguridad	Confío en que el control de acceso por roles protege adecuadamente la información.
Satisfacción	Estoy satisfecho con la plataforma como solución para el monitoreo de modelos de ML.

4. Tabla de Validación de Criterios

La norma ISO/IEC 25010 establece criterios de calidad de software tales como usabilidad, funcionalidad, rendimiento y seguridad para evaluar la calidad de un producto de software. A continuación, se presenta la tabla de validación que incluye preguntas específicas para cada criterio evaluado en la plataforma de monitoreo:

Criterio / Pregunta de Validación	Evaluación (1 a 5)
Usabilidad	
¿La plataforma es intuitiva y fácil de usar para usuarios con distintos roles (analistas, supervisores)?	5
¿La navegación entre los diferentes módulos (métricas, estabilidad, clientes, SHAP) es clara y fluida?	3
¿El diseño visual facilita la lectura e interpretación de los gráficos y resultados?	4
¿El sistema de roles y permisos es fácil de entender para el usuario final?	3
¿La plataforma se adapta correctamente a la resolución y dispositivo del usuario?	3

Rendimiento	
¿La plataforma responde con rapidez al consultar métricas de performance de los modelos?	5
¿Las visualizaciones de estabilidad (PSI, CSI) cargan sin demoras significativas?	5
¿El agente LLM responde en tiempos aceptables ante consultas sobre los modelos?	3
¿El acceso al análisis SHAP e interpretación de variables es ágil?	3
Funcionalidad	
¿Las métricas de performance (Gini, KS, AUC) se muestran correcta y completamente?	5
¿El módulo de análisis de clientes permite filtrar y consultar información de manera efectiva?	5
¿El agente LLM responde de forma precisa y coherente a preguntas sobre los modelos?	4
¿Los gráficos de SHAP representan adecuadamente la contribución de cada variable?	4
¿El control de acceso por roles restringe correctamente los módulos según el perfil del usuario?	3
Seguridad	
¿El acceso a la plataforma mediante usuario y contraseña (desplegada en AWS) es seguro y confiable?	4
¿Los usuarios solo pueden ver la información correspondiente a su rol asignado?	5
¿Confío en que la información sensible de los modelos y clientes está protegida en la plataforma?	5
Satisfacción General	
¿La plataforma cubre las necesidades de monitoreo de modelos de ML del equipo de Model Risk Management?	4
¿Recomendaría el uso de esta plataforma como herramienta de seguimiento de modelos?	4
¿Considera que la plataforma es un aporte valioso para el proceso de validación y monitoreo de modelos?	4
Evalúe de manera general la plataforma de monitoreo de modelos de ML.	4

5. Comentarios del Evaluador (Opcional)

La plataforma Integra todo el flujo de monitoreo y la interpretación , ademas el agente LLM potencia la interaccion.

6. Confirmación de Aprobación

El presente documento ha sido revisado y completado por el evaluador, quien certifica la validación de la Plataforma de Monitoreo de Modelos de Machine Learning.

Fecha:	08 de mayo de 2026
Firma:	

ANEXO H. Validación de la Plataforma Web MRM Suite (Experto 03)

VALIDACIÓN DEL ARTEFACTO

Plataforma de Monitoreo de Modelos de Machine Learning

Información del Evaluador	
Nombres y apellidos:	Bryan Josue Hinostroza Nuñez
Título / Profesión:	Ingeniero de Sistemas
Grado académico:	Ingeniero
Años de experiencia en el campo relacionado con la profesión:	9 años
Área(s) de especialización:	Business Analytics
Cargo en la empresa/organización:	Model Risk Monitoring Associate

1. Introducción

La validación del artefacto de software es un proceso sistemático que verifica si un sistema cumple con los requisitos establecidos y satisface las necesidades de sus usuarios. Este documento presenta la validación de la Plataforma de Monitoreo de Modelos de Machine Learning, desarrollada para el equipo de Model Risk Management.

La plataforma permite visualizar métricas de performance y estabilidad de modelos, analizar clientes, interpretar variables mediante gráficos SHAP, y consultar información a través de un agente LLM integrado. El acceso está restringido por roles y la plataforma se encuentra desplegada en AWS, con ingreso únicamente mediante usuario creado.

El propósito de esta validación es evaluar la calidad de la plataforma desde la perspectiva de los usuarios finales del equipo de Model Risk Management, considerando los criterios de usabilidad, rendimiento, funcionalidad, seguridad y satisfacción general, en concordancia con el estándar ISO/IEC 25010.

2. Descripción del Artefacto

La Plataforma de Monitoreo de Modelos de ML es una aplicación web desplegada en AWS que centraliza el seguimiento y análisis de los modelos de riesgo crediticio en producción. Sus principales módulos son:

- Métricas de Performance: Visualización de indicadores clave como Gini, KS y AUC-ROC por modelo.
- Estabilidad de Modelos: Análisis de PSI (Population Stability Index) y CSI (Characteristic Stability Index) para detectar degradación.
- Análisis de Clientes: Consulta y filtrado de información de clientes vinculados a los modelos.
- Interpretabilidad SHAP: Gráficos de contribución de variables para la explicación de predicciones individuales y globales.
- Agente LLM: Asistente conversacional para consultas sobre métricas, documentación y comportamiento de los modelos.
- Control de Acceso por Roles: Gestión de permisos que restringe el acceso a módulos según el perfil del usuario.

3. Escala de Evaluación

Para la evaluación de la plataforma se utilizó la escala Likert de 5 puntos que se presenta en la siguiente tabla:

Puntaje	Significado
1	Muy Deficiente
2	Deficiente
3	Aceptable
4	Bueno
5	Muy Bueno

Los criterios y preguntas de validación empleados para la evaluación se presentan en la siguiente tabla:

Aspecto	Pregunta de Evaluación
Usabilidad	La plataforma es intuitiva y fácil de navegar para usuarios con distintos roles.
Rendimiento	La plataforma responde con rapidez al consultar métricas, gráficos y análisis de modelos.
Funcionalidad	Las herramientas de monitoreo, el agente LLM y el análisis SHAP funcionan correctamente.
Seguridad	Confío en que el control de acceso por roles protege adecuadamente la información.
Satisfacción	Estoy satisfecho con la plataforma como solución para el monitoreo de modelos de ML.

4. Tabla de Validación de Criterios

La norma ISO/IEC 25010 establece criterios de calidad de software tales como usabilidad, funcionalidad, rendimiento y seguridad para evaluar la calidad de un producto de software. A continuación, se presenta la tabla de validación que incluye preguntas específicas para cada criterio evaluado en la plataforma de monitoreo:

Criterio / Pregunta de Validación	Evaluación (1 a 5)
Usabilidad	
¿La plataforma es intuitiva y fácil de usar para usuarios con distintos roles (analistas, supervisores)?	4
¿La navegación entre los diferentes módulos (métricas, estabilidad, clientes, SHAP) es clara y fluida?	4
¿El diseño visual facilita la lectura e interpretación de los gráficos y resultados?	4
¿El sistema de roles y permisos es fácil de entender para el usuario final?	5
¿La plataforma se adapta correctamente a la resolución y dispositivo del usuario?	4

Rendimiento	
¿La plataforma responde con rapidez al consultar métricas de performance de los modelos?	4
¿Las visualizaciones de estabilidad (PSI, CSI) cargan sin demoras significativas?	5
¿El agente LLM responde en tiempos aceptables ante consultas sobre los modelos?	4
¿El acceso al análisis SHAP e interpretación de variables es ágil?	4
Funcionalidad	
¿Las métricas de performance (Gini, KS, AUC) se muestran correcta y completamente?	4
¿El módulo de análisis de clientes permite filtrar y consultar información de manera efectiva?	5
¿El agente LLM responde de forma precisa y coherente a preguntas sobre los modelos?	5
¿Los gráficos de SHAP representan adecuadamente la contribución de cada variable?	4
¿El control de acceso por roles restringe correctamente los módulos según el perfil del usuario?	5
Seguridad	
¿El acceso a la plataforma mediante usuario y contraseña (desplegada en AWS) es seguro y confiable?	5
¿Los usuarios solo pueden ver la información correspondiente a su rol asignado?	5
¿Confío en que la información sensible de los modelos y clientes está protegida en la plataforma?	5
Satisfacción General	
¿La plataforma cubre las necesidades de monitoreo de modelos de ML del equipo de Model Risk Management?	5
¿Recomendaría el uso de esta plataforma como herramienta de seguimiento de modelos?	5
¿Considera que la plataforma es un aporte valioso para el proceso de validación y monitoreo de modelos?	5
Evalúe de manera general la plataforma de monitoreo de modelos de ML.	5

5. Comentarios del Evaluador (Opcional)

6. Confirmación de Aprobación

El presente documento ha sido revisado y completado por el evaluador, quien certifica la validación de la Plataforma de Monitoreo de Modelos de Machine Learning.

Fecha:	10 de mayo de 2026
Firma:	