

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

CÁLCULO DE LA PROBABILIDAD DE INCUMPLIMIENTO PARA LA
IMPLEMENTACIÓN DENTRO DEL PROCESO DE SEGUIMIENTO DEL NIVEL DE
RIESGO DE UNA CARTERA DE CRÉDITO POR MEDIO DE UN SCORING DE
COMPORTAMIENTO

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del
Programa de Posgrado en Estadística para optar al grado y título de Maestría Profesional en
Estadística

YESI WILMER GONZÁLEZ MÉNDEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica

DEDICATORIA Y AGRADECIMIENTOS

Este trabajo está dedicado a mi madre que me inculcó el amor por el estudio y desde pequeño me dio el impulso y apoyo para alcanzar mis objetivos. A través de sus enseñanzas, se imprimió una huella imborrable en mi vida, y mi éxito académico es un reflejo de su fe inquebrantable.

Agradezco a mi tutor MSc Esteban Bermúdez Aguilar por el apoyo brindado durante el proyecto, por toda su disposición a colaborar, por su tiempo y excelente asesoramiento en el desarrollo de este trabajo.

También agradezco a mis lectores, MSc Max Alberto Rimola Guevara, y Ph.D. Shu Wei Chou Chen, por aceptar ser parte de este proyecto y por la colaboración brindada.

Doy gracias, al alma máter, Universidad de Costa Rica, por el aprendizaje que coseché durante mi estadía en la institución y particularmente a la Escuela de Estadística por brindarme educación de calidad y convertirme en un profesional integral y competente.

Así mismo, agradezco a la Junta de Pensiones y Jubilaciones del Magisterio Nacional, por la oportunidad de realizar este estudio para proporcionar información útil para la toma de decisiones en pro del bienestar de sus afiliados.

Por último, doy gracias a mi madre, profesores, amigos y todas las personas que de alguna u otra manera mostraron su apoyo, preocupación e interés por la culminación con éxito de este proyecto. Les agradezco toda la paciencia que me tuvieron en los momentos de estrés y cada frase motivadora que me extendieron.

Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística.

M.Sc. Victoria Hernández Mora
**Representante de la Decana
Sistema de Estudios de Posgrado**

M.Sc. Esteban Bermúdez Aguilar
Profesor Guía

M.Sc. Max Alberto Rimola Guevara
Lector

Ph.D. Shu Wei Chou Chen
Lector

Ph.D. Gilbert Brenes Camacho
Director Programa de Posgrado en Estadística

Yesi Wilmer González Méndez
Sustentante

TABLA DE CONTENIDO

DEDICATORIA Y AGRADECIMIENTOS.....	ii
HOJA DE APROBACION.....	iii
TABLA DE CONTENIDO	iv
RESUMEN	vi
LISTA DE CUADROS.....	vii
LISTA DE GRÁFICOS.....	ix
LISTA DE FIGURAS.....	xi
LISTA DE ABREVIATURAS.....	xii
CAPÍTULO I. INTRODUCCIÓN.....	1
CAPÍTULO II. ESTADO DE LA CUESTION.....	5
2.1. ORIGEN DE LA JUNTA DE PENSIONES DEL MAGISTERIO NACIONAL	5
2.2. DESCRIPCIÓN DEL PORTAFOLIO DE INVERSIONES DEL RCC	7
2.3. DESCRIPCIÓN DE LA CARTERA DE CRÉDITOS DEL RCC	8
2.4. GESTIÓN DE RIESGO DE CRÉDITO SOBRE LA CARTERA DEL RCC ...	10
2.5. PROBLEMÁTICA PARA RESOLVER	16
2.6. REFERENTES O ELEMENTOS TEÓRICOS	17
CAPÍTULO III. METODOLOGÍA.....	25
3.1. FUENTE DE DATOS	25
3.2. DEFINICIÓN DE VARIABLES DE ESTUDIO	26
3.3. EVIDENCIAS DE CALIDAD DE LA MEDICIÓN PARA LAS VARIABLES DEL ESTUDIO	29
3.4. MODELOS ESTADÍSTICOS DE ANÁLISIS POR EMPLEAR	30
3.5. GENERACIÓN DEL PLAN DE PRUEBAS	43
CAPITULO IV. RESULTADOS	49

4.1. ANÁLISIS DESCRIPTIVO	49
4.2. MODELOS DE APRENDIZAJE NO SUPERVISADO	55
4.3. MODELOS DE APRENDIZAJE SUPERVISADO	59
4.4. VALIDACIÓN DE RESULTADOS Y CONSTRUCCIÓN DE CATEGORÍAS DE RIESGO	64
4.5. RESULTADOS DE CÁLCULO DE PERDIDA ESPERADA	70
CAPÍTULO V. CONCLUSIONES	74
BIBLIOGRAFÍA	78
ANEXOS	85

RESUMEN

El propósito de la investigación es modelar la probabilidad de incumplimiento de las operaciones crediticias de la cartera de créditos del Fondo de Pensiones del Régimen de Capitalización Colectiva, que administra la Junta de Pensiones y Jubilaciones del Magisterio Nacional, como una oportunidad de innovación respecto a esquemas comúnmente utilizados por entidades financieras, y que a su vez aporte a maximizar los rendimientos de la cartera.

Para el abordaje de este objetivo se lleva a cabo un ejercicio de minería de datos y la aplicación de modelos de aprendizaje de máquinas, que permita definir un modelo para la construcción de un *score* de crédito de seguimiento, que arroje la probabilidad de incumplimiento por operación de crédito. Con este insumo, se procede a construir un esquema de calificación de riesgo del deudor que sirva como insumo para el cálculo de pérdidas crediticias esperadas. Para lo anterior, se utiliza información sobre características socioeconómicas de los deudores para el año 2023 y 2024, siendo utilizados de manera enmascarada con el propósito de resguardar la respectiva confidencialidad de la información.

Dentro de las principales conclusiones, de un total de 28 variables iniciales se desarrolla un proceso de selección de variables basado en análisis descriptivo y multivariado, que permite definir un total de cinco variables explicativas, referentes a la morosidad al cierre de cada mes y de manera histórica, así como indicadores sobre la situación laboral del deudor. La variable respuesta, categoriza a las operaciones entre pertenecientes a un buen y mal pagador.

La comparación y evaluación de resultados de una serie de modelos de clasificación siguiendo un proceso de minería de datos permite llegar a la conclusión de que el modelo de clasificación *Naive Bayes* arroja los mejores indicadores de clasificación, con una clasificación correcta global de 95,8%, de buenos pagadores de 96,3% y de malos pagadores de 82,0%. Luego se construye un esquema de calificación de riesgo a partir de las probabilidades predichas, alcanzando resultados razonables.

Finalmente, al llevar este resultado sobre el proceso de cálculo de pérdida esperada de la cartera, se obtiene que el monto de pérdida esperada recalculado es en promedio ¢5.076 millones más bajo, lo cual se ubica dentro del nivel de apetito al riesgo definido institucionalmente, y se concluye que el resultado es viable y aporta a maximizar los rendimientos de la cartera.

LISTA DE CUADROS

Cuadro 1. Criterios de asignación de calificaciones de riesgo según indicadores de riesgo de crédito, diciembre 2024	11
Cuadro 2. Distribución de cartera de crédito del RCC por cantidad de operaciones según calificación de riesgo deudor, diciembre 2024	12
Cuadro 3. Probabilidades de Incumplimiento por matrices de transición según calificación de riesgo deudor, diciembre 2024.....	14
Cuadro 4. Representación de matriz de confusión	44
Cuadro 5. Parámetros evaluados.....	47
Cuadro 6. Distribución de las variables categóricas según buen pagador, enero y diciembre 2023	49
Cuadro 7. Promedio de variables continuas según buen y mal pagador, enero y diciembre 2023	52
Cuadro 8. Promedio de variables continuas moderadamente relevantes por condición laboral del deudor según buen pagador, enero y diciembre 2023.....	54
Cuadro 9. Matriz de correlaciones variables de morosidad, enero y diciembre 2023	56
Cuadro 10. Análisis ACP variables de morosidad, enero y diciembre 2023.....	56
Cuadro 11. Cargas factoriales ACP variables de morosidad, enero y diciembre 2023	57
Cuadro 12. Centros de clúster K medias para variables de morosidad, enero y diciembre de 2023	58
Cuadro 13. Clúster por K medias para variables de morosidad según buen pagador, enero y diciembre de 2023.....	58
Cuadro 14. Selección de variables para fase de modelado	59
Cuadro 15. Resultados de evaluación de modelos calibrados, enero y diciembre de 2023..	62
Cuadro 16. Distribución de probabilidades predichas sobre base de datos nuevos, enero a diciembre de 2023.....	65
Cuadro 17. Distribución probabilidades predichas según cuartiles, enero a diciembre de 2023	65
Cuadro 18. Definición de categorías de riesgo según probabilidades de incumplimiento, enero a diciembre de 2023	66

Cuadro 19. Distribución de operaciones en datos de validación según categorías de riesgo construidas, enero a diciembre de 2023.....	67
Cuadro 20. Distribución de variables cualitativas según categorías de riesgo, enero a diciembre de 2023.....	68
Cuadro 21. Resultados promedio de variables cuantitativas según categorías de riesgo, enero a diciembre de 2023.....	69
Cuadro 22. Resultados y comparación de pérdida esperada de la cartera por mes, enero a diciembre de 2024.....	71
Cuadro 23. Comparación indicadores de cobertura registrada y calculada respecto a saldo de cartera con más de 90 días de atraso, enero - diciembre 2024.....	73
Cuadro 24. Indicadores de clasificación alternando variables de mora máxima y grupo de mora según métodos de clasificación, enero y diciembre 2023	85
Cuadro 25. Indicadores de clasificación alternando parámetros según métodos de clasificación, enero y diciembre 2023	86
Cuadro 26. Indicadores de clasificación alternando parámetros según métodos de clasificación, enero y diciembre 2023	87

LISTA DE GRÁFICOS

Gráfico 1. Distribución de portafolio de inversiones del RCC, diciembre 2024.....	8
Gráfico 2. Indicadores de morosidad cartera del RCC, enero - diciembre 2024.....	9
Gráfico 3. Monto de Perdida Esperada de la cartera de crédito del RCC, enero - diciembre de 2024	15
Gráfico 4. Indicador KRI de cobertura de PE respecto a mora con más de 90 días de atraso y límite de apetito al riesgo del RCC, enero - diciembre 2024.....	16
Gráfico 5. Circulo de correlación del ACP sobre variables de morosidad, enero y diciembre de 2023.....	57
Gráfico 6. Representación de Curva ROC según modelo Bayesiano.....	63
Gráfico 7. Representación de Curva AUC PR según modelo Bayesiano.....	64
Gráfico 8. Comparación monto de perdida esperada registrada y calculada con indicadores de mora de la cartera, enero - diciembre 2024.....	72
Gráfico 9. Distribución de días de mora de operaciones de crédito según categorías de riesgo, febrero - noviembre 2023	88
Gráfico 10. Distribución de mora promedio del deudor según categorías de riesgo, febrero - noviembre 2023	88
Gráfico 11. Distribución de puntaje histórico deudor según categorías de riesgo, febrero - noviembre 2023	89
Gráfico 12. Distribución de segmento de mercado según categorías de riesgo, febrero - noviembre 2023	89
Gráfico 13. Distribución de tenencia de operaciones incobrables según categorías de riesgo, febrero - noviembre 2023	90
Gráfico 14. Distribución de días de mora de operaciones de crédito según categorías de riesgo, enero - diciembre 2024	90
Gráfico 15. Distribución de mora promedio del deudor según categorías de riesgo, enero - diciembre 2024	91
Gráfico 16. Distribución de puntaje histórico deudor según categorías de riesgo, enero - diciembre 2024	91
Gráfico 17. Distribución de segmento de mercado según categorías de riesgo, enero - diciembre 2024	92

Gráfico 18. Distribución de tenencia de operaciones incobrables según categorías de riesgo,
enero - diciembre 202492

LISTA DE FIGURAS

Figura 1. Tipos de score de crédito	2
Figura 2. Diagrama de validación cruzada usando K grupos	47

LISTA DE ABREVIATURAS

CONASSIF	Consejo Nacional de Supervisión del Sistema Financiero
INA	Instituto Nacional de Aprendizaje
ISO	Del inglés International Organization for Standardization
JUPEMA	Junta de Pensiones y Jubilaciones del Magisterio Nacional
MEP	Ministerio de Educación Pública
RCC	Régimen de Capitalización Colectiva
RTR	Régimen Transitorio de Reparto
SUGEF	Superintendencia General de Entidades Financieras
SUPEN	Superintendencia de Pensiones

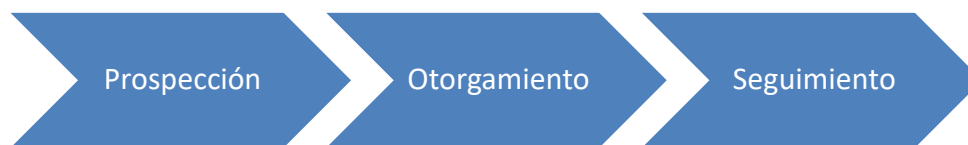
CAPÍTULO I. INTRODUCCIÓN

Según la definición del Consejo Nacional de Supervisión del Sistema Financiero (CONASSIF) contenida en el acuerdo SUGEF 2-10, la administración integral de riesgos se entiende como un proceso por medio del cual una entidad financiera identifica, mide, evalúa, monitorea, controla, mitiga y comunica los distintos tipos de riesgo a que se encuentra expuesta, entre ellos el más relevante es el riesgo de crédito. Por otro lado, en la terminología establecida por la ISO 31000 sobre Administración de Riesgo, en la definición 3.7 del documento, la palabra “probabilidad” se utiliza para indicar la posibilidad de que algo suceda, esté definida, medida o determinada objetiva o subjetivamente, cualitativa o cuantitativamente, y descrita utilizando términos generales o matemático.

En este sentido, la probabilidad de incumplimiento de un deudor es un indicador que se otorga a un cliente dueño de una operación de crédito, con el fin de cuantificar el riesgo asociado a la posibilidad de que no se honren las obligaciones crediticias adquiridas en un momento dado. En el contexto bancario una herramienta que permite generar este indicador es lo que se conoce como *scoring* de crédito.

Según muestra la figura 1, existen diferentes tipos de *scoring* de crédito, cuya definición se determina según la ubicación dentro del proceso de otorgamiento de créditos de una institución. Se tienen los *scoring* prospectivos que buscan identificar con datos básicos si una determinada persona puede ser un cliente potencial, luego se tienen los *scoring* de otorgamiento, que se utilizan para estimar si a un cliente potencial se le puede otorgar un crédito en función de su situación socioeconómica y de la deuda que se desea asumir, finalmente se tienen los *scoring* de seguimiento, que tienen la función de monitorear, una vez otorgado el crédito, la probabilidad de que una operación pueda caer en impago, esto en la medida de que cambie en el tiempo la situación socioeconómica del deudor.

Figura 1.
Tipos de score de crédito



Fuente: Grupo Financiero BBVA

Para la construcción de esta herramienta, se suelen utilizar modelos estadísticos multivariados, que buscan modelar la probabilidad de incumplimiento del deudor de una manera objetiva y eficiente, por lo que el uso de herramientas de *score* crediticio funciona como un elemento fundamental que permiten agilizar procesos, eliminar el juicio subjetivo, y medir de mejor manera el nivel de riesgo de los deudores, con el propósito de mejorar la calidad crediticia, y por ende la rentabilidad de una cartera de crédito.

Por su parte, el uso de la probabilidad de incumplimiento funciona como indicador de riesgo de crédito de diferentes formas, desde decidir si un crédito se otorga a un determinado deudor, hasta ser utilizado como insumo para la definición de montos de pérdida esperada por eventuales impagos.

De esta manera, la presente propuesta busca desarrollarse con datos pertenecientes a la Junta de Pensiones del Magisterio Nacional (JUPEMA), una entidad pública no estatal, con personería jurídica y patrimonio propio, que está sujeta a normas de Ley y al ordenamiento jurídico administrativo-público y se encuentra bajo la fiscalización de la Superintendencia de Pensiones (SUPEN), regulada por la ley 7531 y sus reformas.

JUPEMA tiene a cargo la administración del fondo del Régimen de Capitalización Colectiva (RCC), el cual es uno de los fondos creados para cubrir el retiro de los trabajadores de la educación costarricense y recibe aportes patronales, obreros y estatales para dicho fin según lo establecido en el artículo 104 de la ley 7531 y sus reformas. Tiene como misión garantizar una pensión digna para los trabajadores de la educación, por lo que es imperativo maximizar la capacidad del fondo para hacer frente al pago de pensiones, lo cual se logra por medio de

la ejecución de inversión de las cotizaciones que recibe el fondo, para generar rendimientos que colaboren a obtener un mayor crecimiento de este y mejorar su sostenibilidad en el tiempo.

Dentro de las actividades que tiene JUPEMA para invertir y generar rendimiento, está la inversión en el mercado de valores de nacional e internacional y la colocación de recursos en créditos hipotecarios y de consumo a sus afiliados; sin embargo, al ser esta última una actividad con mayor riesgo, se espera un mayor rendimiento que lo compense respecto al mercado bursátil, por lo que es primordial que cuente con una cartera sana para lograr el objetivo que persigue, pero sin asumir riesgos elevados.

Sobre el nivel de riesgo asumido, este se monitorea por medio de distintos indicadores, entre ellos se tienen indicadores de morosidad, cálculo y seguimiento de categorías de riesgo del deudor, cálculo y cobertura de estimaciones por pérdidas crediticias esperadas, y cálculo de probabilidades de incumplimiento según análisis de matrices de transición.

Como proceso de mejora, actualmente se están analizando distintas opciones metodológicas para robustecer los modelos e indicadores empleados, entre ellos, la metodología de calificación de riesgo del deudor y la probabilidad de incumplimiento derivada, esto debido a que diferentes estudios han mostrado que las metodologías actuales presentan deficiencias que podrían estar sobreestimando el nivel de riesgo de la cartera, por lo que surge la posibilidad de explorar métodos alternativos que disminuyan los sesgos observados.

De esta manera, para la presente propuesta, se busca desarrollar un *score* crediticio de seguimiento para la cartera de crédito del RCC administrada por JUPEMA, para cumplir con dos objetivos, generar un esquema metodológico por medio de modelos estadísticos multivariados que permitan modelar de mejor manera la probabilidad de incumplimiento de los deudores, y de esta manera, utilizar este insumo dentro del proceso de generación de estimaciones o pérdidas crediticias esperadas, siempre que el resultado se ubique dentro del apetito al riesgo que la institución desea asumir.

Lo anterior representa dos oportunidades importantes de innovación respecto al esquema actual utilizado, por un lado, aporta un nuevo método de clasificación de los deudores

utilizando esquemas más robustos e integrales, basados en los diferentes perfiles de los deudores y su situación crediticia, y por otra parte este insumo permite estimar de mejor manera las estimaciones crediticias por pérdidas esperadas, lo que favorece a maximizar el rendimiento de la cartera.

Específicamente se establecen los siguientes objetivos

Objetivo general:

- Generar un *score* crediticio de seguimiento para la cartera de crédito del fondo de pensiones del Régimen de Capitalización Colectiva de Costa Rica, con el propósito de calcular la probabilidad de incumplimiento por operación crediticia y definir un esquema de calificación de riesgo que sirva como insumo para el cálculo de pérdidas crediticias esperadas y de esta manera aportar en maximizar la rentabilidad de la cartera para el año 2024.

Objetivos específicos

- Realizar un proceso de selección de variables por medio de un análisis descriptivo y multivariado para identificar las variables con mayor poder predictivo.
- Evaluar un modelo de aprendizaje supervisado adecuado, con el propósito de generar probabilidades de incumplimiento por operación crediticia, utilizando técnicas de aprendizaje de máquinas con validación cruzada, medidas de desempeño y validación con datos fuera de tiempo, que permita definir un esquema de calificación de riesgo de la operación crediticia para el cálculo de pérdidas esperadas de la cartera.
- Determinar si el resultado obtenido representa una alternativa viable como insumo dentro del proceso de cálculo de pérdidas crediticias esperadas.

En los siguientes capítulos se desarrollará el contexto actual que motiva el estudio, la definición del esquema metodológico a seguir, y posteriormente se presentan los principales resultados y conclusiones.

CAPÍTULO II. ESTADO DE LA CUESTION

2.1. ORIGEN DE LA JUNTA DE PENSIONES DEL MAGISTERIO NACIONAL

Según como comenta Robles (2014), la creación de los regímenes de pensiones en Costa Rica está ligada al desarrollo del sistema educativo del país, ya que se tiene que la primera ley de pensiones se establece con la Ley General de Educación Común de 1886, la cual creó el primer fondo de pensiones del país para los docentes públicos. Luego, en 1935 se aprueba la Ley General de Pensiones, con la cual se extienden los beneficios de pensión a otros grupos de la población, muchas de ellas de tipo no contributivo y se establece un monto de pensión mínima.

La Junta de Pensiones y Jubilaciones del Magisterio Nacional (JUPEMA) nace el 5 de setiembre de 1958 a través de la creación de la Ley de Pensiones y Jubilaciones del Magisterio Nacional 2248, como una institución que busca brindar a los trabajadores de la educación costarricense la cobertura sobre los riesgos de invalidez, vejez y muerte, por medio de un régimen de pensiones propio.

Tal como explica Robles (2014), el primer régimen de pensiones de Magisterio Nacional inició con aportes que ingresaban directamente a las arcas del Estado, pero nunca se creó un fondo para el pago de pensiones, por lo que el Gobierno se encargó de hacer los pagos de los beneficios con cargo al presupuesto estatal. En sus inicios, se dio un sobrante, ya que se tenía un flujo importante de aportes y poco gasto de pensiones, por lo que se utilizó para el financiamiento estatal, sin embargo, para mediados de la década de los ochenta, el efecto del pago de las pensiones del Magisterio sobre el gasto público creció de manera significativa.

La situación llevo a que el 14 de noviembre de 1991 se aprobara la Ley 7268 Reforma Integral a la Ley de Pensiones y Jubilaciones del Magisterio Nacional, con el objetivo de reformar las condiciones de pensión y los derechos adquiridos, sin embargo, dicha reforma no fue suficiente, por lo que en el año 1995 se establece una nueva reforma con la Ley Reforma Integral de Sistema de Pensiones y Jubilaciones del Magisterio 7531.

A partir de las reformas planteadas, se definen dos regímenes separados, el Régimen Transitorio de Reparto (RTR), donde se incluyen todas aquellas personas que empezaron a laborar en el sector educación antes del 15 de julio de 1992, se le llama transitorio porque

tiende a desaparecer en el tiempo pues es un régimen cerrado que no admite nuevos cotizantes. Por su parte se crea el Régimen de Capitalización Colectiva (RCC) para todos los trabajadores de la educación que empezaron a laborar en el Magisterio Nacional después del 15 de julio de 1992, con una estructura de beneficios diferenciada.

De esta manera, según el artículo 104 de la Ley 7531, le corresponde a JUPEMA la administración total del RCC, desde la recaudación de cotizaciones, administración del fondo, inversión de las cotizaciones, hasta el otorgamiento del derecho de pensión y el pago de estas. Por su parte, a JUPEMA le corresponde la administración parcial del RTR, lo que incluye únicamente la recaudación de las cotizaciones, y el otorgamiento del derecho de pensión, las demás funciones las ejecuta el estado, por lo que el pago de pensiones de este régimen lo cubre el presupuesto nacional. Se resalta que JUPEMA es una institución con autonomía en sus decisiones, pero supervisada y fiscalizada por la Superintendencia de Pensiones (SUPEN) y la Contraloría General de la Republica.

Según la información publicada por JUPEMA en su página electrónica, el principal objetivo de JUPEMA es garantizar una pensión digna para los trabajadores de la educación costarricense, por lo que se está en constante búsqueda de maximizar el crecimiento del fondo, desde el resguardo de que se honre el pago de las cotizaciones de sus afiliados, hasta maximizar los rendimientos de las inversiones por medio de un adecuado Gobierno Corporativo. De esta manera se establecen la misión y visión de la institución:

- Misión: Ser la institución de Seguridad Social responsable de la gestión eficiente del fondo de pensiones de nuestros afiliados, para mejorar su calidad de vida.
- Visión: Ser el referente de seguridad social costarricense en la administración sostenible e innovadora de los sistemas de pensiones.

Son afiliados a JUPEMA los trabajadores activos y pensionados del magisterio nacional, adscritos tanto al RCC como al RTR. En cuanto a trabajadores activos incluye a los funcionarios y educadores del Ministerio de Educación Pública (MEP), así como de las Universidades Públicas¹, y del Instituto Nacional de Aprendizaje (INA), junto con los

¹ Universidad de Costa Rica (UCR), Universidad Nacional (UNA), Instituto tecnológico de Costa Rica (TEC), y Universidad Técnica Nacional (UTN), Universidad Estatal a Distancia (UNED).

funcionarios y educadores de Centros Educativos Privados, que operen en el país, tanto de educación preescolar como primer y segundo ciclo. Lo anterior representa, al cierre de 2024, una base de afiliados de más de 166 mil personas.

Para diciembre de 2024, el RCC alcanzó un monto de ¢5,5 billones, donde según sus estados financieros con corte a diciembre 2024, alrededor de 50% del fondo se ha generado con los rendimientos obtenidos de sus inversiones, y tiene solvencia para garantizar el pago de pensiones actuales y futuras, esto según lo mostrado en el Informe Anual 2024 publicado por JUPEMA.

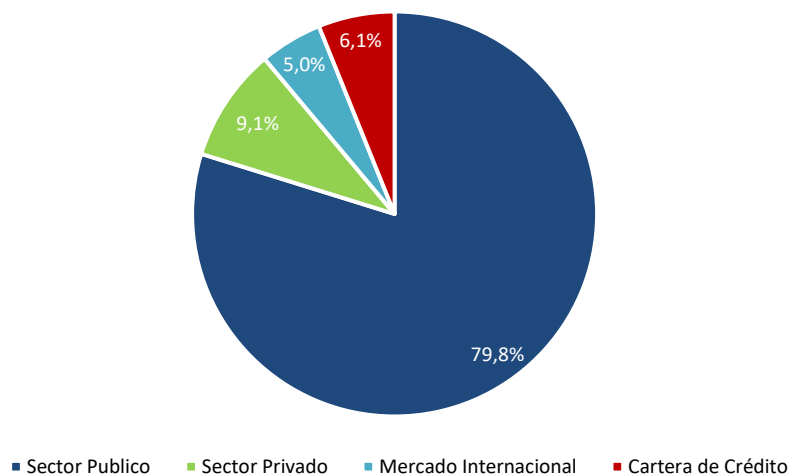
2.2. DESCRIPCIÓN DEL PORTAFOLIO DE INVERSIONES DEL RCC

En cuanto a las posibilidades de inversión que tiene JUPEMA para el crecimiento del fondo de pensiones del RCC, la Ley 7531 y sus reformas establecen en el artículo 21 lo siguiente:

- Invertir al menos 30% del Fondo, en valores emitidos por el sector público.
- Invertir hasta un máximo de 20% de la totalidad del Fondo en valores e instrumentos transados en mercados extranjeros debidamente regulados y supervisados.
- Invertir hasta un máximo de 20% en proyectos de infraestructura pública o privada en el territorio nacional, de conformidad con la legislación nacional y lo que al efecto establezca la normativa emitida por el Consejo Nacional de Supervisión del Sistema Financiero y la Superintendencia de Pensiones.
- Invertir hasta un máximo de 20% en préstamos personales y de vivienda para los afiliados. Los préstamos, en el caso de los afiliados activos deberán cobrarse por deducción salarial; para jubilados o pensionados se deducirá mensualmente de los montos de pensión, las amortizaciones, los intereses y otros cargos.

Según lo anterior, el Grafico 1 muestra la distribución del portafolio de inversiones a diciembre de 2024.

Gráfico 1.
Distribución de portafolio de inversiones del RCC, diciembre 2024



Fuente: JUPEMA, Reporte de Gestión 2024

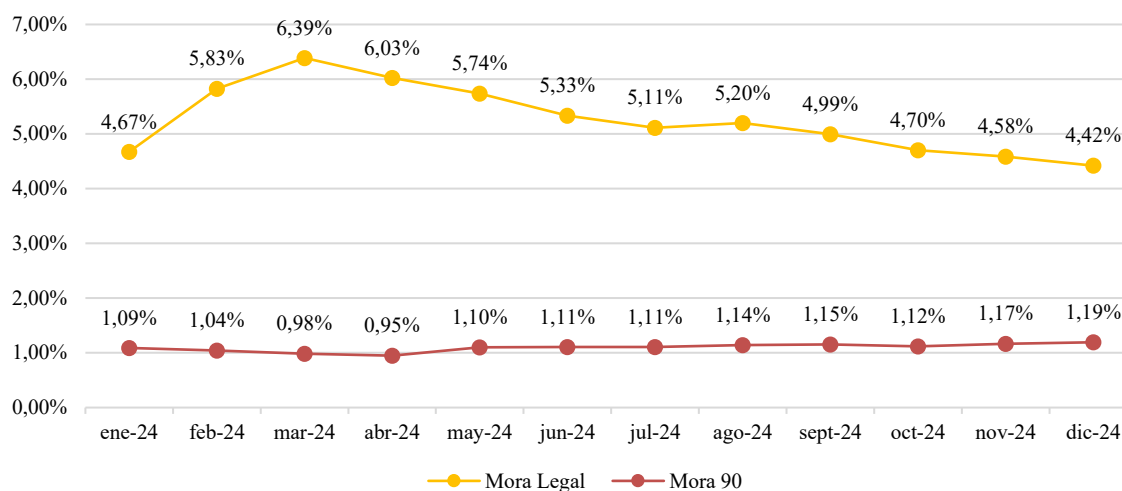
Se observa que el portafolio se concentra principalmente en instrumentos del sector público con 79,8% del total, por su parte, el sector privado nacional concentra 9,1% del portafolio y 5,0% se encuentra invertido en instrumentos del mercado internacional, mientras que 6,1% del portafolio corresponde a la cartera de crédito a los afiliados.

2.3. DESCRIPCIÓN DE LA CARTERA DE CRÉDITOS DEL RCC

En lo referente a la cartera de crédito, alcanza un 6,1% del portafolio de inversiones, lo que representa un monto de ₡327,4 mil millones al cierre de diciembre de 2024, está compuesta por 34.462 operaciones de crédito (todas en colones) con 24.710 deudores.

En cuanto a los indicadores de morosidad, el Gráfico 2 muestra el comportamiento de los indicadores de mora legal (mora a partir de un día de atraso), y mora mayor a 90 días de atraso.

Gráfico 2.
Indicadores de morosidad cartera del RCC, enero - diciembre 2024



Fuente: JUPEMA, Evaluación Actuarial 2024

Se tiene que la mora legal de la cartera mantiene una tendencia de disminución durante el año 2024, pasando de un nivel máximo de 6,39% en marzo a 4,42% en diciembre. Por su parte, la mora mayor a 90 días de atraso se ha mantenido con una tendencia de aumento, pero en niveles relativamente bajos, pasando de 1,09% en enero a 1,19% para el cierre del año.

Otro dato para destacar es que, según artículo 21 de la ley 7531 y sus reformas, el modelo de negocio del programa de crédito de JUPEMA se basa en un mercado cerrado, ya que únicamente se pueden prestar créditos a la población afiliada, y con la característica de que las obligaciones de los deudores deben realizarse por deducción de planilla, lo que favorece la posibilidad de mantener los indicadores de mora relativamente controlados.

Pese a lo anterior, las principales causas de mora se explican por factores de sobreendeudamiento, específicamente con otras entidades que también cuentan con la posibilidad de realizar deducción de planilla, y más recientemente por el efecto que generó la Ley contra la Usura Crediticia en Costa Rica (Ley 9859) que fue aprobada en 2020, o mejor conocida como “Ley de Usura”, la cual incorporó el manejo del salario mínimo inembargable como un tope para la deducción de créditos por planilla para trabajadores activos, lo cual ha provocado limitaciones al proceso de deducciones de créditos. Además, otros factores que afectan el nivel de mora pueden ser las incapacidades, la falta de

nombramientos en deudores con condición laboral interina, la reducción de recargos salariales, entre otros.

2.4. GESTIÓN DE RIESGO DE CRÉDITO SOBRE LA CARTERA DEL RCC

JUPEMA tiene definido un esquema de gestión del riesgo institucional, el cual sigue los lineamientos establecidos en el Reglamento de Gestión de Activos de la SUPEN en su Artículo 13, donde se define la obligación de realizar una declaración de apetito al riesgo institucional con el propósito de declarar los diferentes tipos de riesgo a los que está expuesta la institución según sus objetivos estratégicos, y definir los mecanismos de gestión del riesgo.

De esta manera, según la Declaración de Apetito al Riesgo Institucional publicada en la página web de JUPEMA, se establece que el fondo del RCC está expuesto a riesgo de crédito, tanto por las inversiones en instrumentos de deuda, como a nivel de la cartera de crédito del RCC, por lo que, según lo establecido en el artículo 36 del Reglamento de Gestión de Activos, se debe contar con una metodología de calificación de riesgo del deudor con el propósito de dar seguimiento al nivel de riesgo de la cartera según el deterioro de las condiciones de pago de los deudores.

Para el año 2024 la metodología de calificación de riesgo del deudor vigente para la cartera de crédito del RCC está definida según lineamientos internos, pero tomando elementos de la norma SUGEF 1-05², por lo que está definida por una combinación de 3 indicadores de riesgo (categóricos ordinales) de la siguiente manera:

- **Morosidad promedio:** se refiere a la morosidad promedio del deudor al cierre de mes en sus operaciones de crédito activas, se categoriza en 5 niveles según franjas de cada 30 días de atraso.
- **Comportamiento de pago histórico:** se refiere a un indicador calculado sobre el comportamiento de pago histórico del deudor en sus operaciones de crédito activas en el RCC en los últimos 4 años respecto al momento de la medición, se categoriza en 3 niveles de menor a mayor nivel de riesgo.

² Actualmente derogada, donde se sustituye por el Acuerdo CONASSIF 14-21.

- **Capacidad de pago:** se refiere a la relación entre la suma de las obligaciones de crédito del deudor, y el salario líquido del deudor, la razón resultante se categoriza en 4 niveles de menor a mayor nivel de riesgo.

Según lo anterior, JUPEMA define un total de ocho calificaciones de riesgo en las cuales debe calificar individualmente a los deudores. Éstas se identifican, de menor a mayor nivel de riesgo de crédito siguiendo un patrón alpha numérico, tomando las denominaciones: A1, A2, B1, B2, C1, C2, D y E. Los criterios para asignar las categorías de riesgo según el nivel de variación de los indicadores se muestran en el Cuadro 1.

Cuadro 1.
Criterios de asignación de calificaciones de riesgo según indicadores de riesgo de crédito, diciembre 2024

Calificación de Riesgo	Morosidad	Comportamiento de Pago Histórico	Capacidad de Pago
A1	Nivel 1	Nivel 1	Nivel 1
A2	Nivel 2	Nivel 2	Nivel 1
B1	Nivel 2	Nivel 2	Nivel 2
B2	Nivel 2	Nivel 2	Nivel 2
C1	Nivel 3	Nivel 1	Nivel 3
C2	Nivel 3	Nivel 2	Nivel 3
D	Nivel 4	NA	Nivel 4
E	Nivel 5	Nivel 3	NA

Fuente: JUPEMA

La distribución de cartera según categoría de riesgo por cantidad de operaciones se muestra en el Cuadro 2.

Cuadro 2.
Distribución de cartera de crédito del RCC por cantidad de operaciones según calificación de riesgo deudor, diciembre 2024

Calificación de Riesgo	Operaciones	Porcentaje
A1	30 873	89,6%
A2	299	0,9%
B1	1 053	3,1%
B2	48	0,1%
C1	119	0,3%
C2	36	0,1%
D	320	0,9%
E	1 714	5,0%
Total	34 462	100,0%

Fuente: JUPEMA

Se obtiene que 89,6% de la cartera se ubica en categoría de riesgo A1, en menor medida se observa una concentración de cartera entre 0,1% y 3,1% en las restantes categorías de riesgo desde A2 hasta la categoría D, mientras que la categoría de mayor riesgo, la categoría E, alcanza una concentración de 5,0% del total de operaciones al cierre de 2024.

Seguidamente, según el artículo 42 del mismo reglamento se establece que:

“Conforme la calificación de riesgo asignada a cada deudor, la entidad debe definir la metodología para el cálculo de las estimaciones necesarias y suficientes para cubrir las posibles pérdidas ante la falta de pago por parte del deudor de los créditos concedidos, la cual debe estar registrada al cierre de cada mes.

Esta metodología debe considerar para el cálculo de la estimación, el valor ajustado de las garantías como un mitigador de riesgo; el cual corresponde a la valoración de la garantía por el porcentaje de su aceptación.

En caso de que un crédito se determine como incobrable, la entidad regulada debe estimar el cien por ciento del saldo adeudado.”

Según lo anterior, JUPEMA tiene definida una metodología interna que recoge los elementos mencionados, junto con consideraciones descritos en las Normas Internacionales de Información Financiera número 9 (NIIF 9), de esta manera las estimaciones se definen como

la Perdida Esperada (PE) de la cartera de crédito por eventuales impagos, y para obtener el resultado de la perdida esperada de la cartera se parte de la siguiente definición:

$$PE = \sum_{i=1}^n PI_i * EAD_i * LGD_i \quad [1]$$

Donde:

- PE: Perdida Esperada de la cartera
- PI: Probabilidad de Incumplimiento de la operación i, definida según categoría de riesgo del deudor.
- EAD: Exposición o perdida potencial de la operación de crédito i, se refiere al saldo a valor presente del principal de la operación de crédito.
- LGD: Perdida dado el incumplimiento de la operación de crédito i, lo cual considera el valor ajustado de las garantías como mitigador de riesgo de crédito.

En este sentido, los valores correspondientes a los componentes de EAD y de LGD están dados por las características propias de la operación, pero los valores de PI deben estar alineados a la metodología de calificación de riesgo del deudor.

Según el Acuerdo CONASSIF 14-21, para el cálculo de las probabilidades de incumplimiento de las operaciones de crédito, JUPEMA tiene la posibilidad de contar con una definición propia, por lo que tiene establecido que dichas probabilidades surgen del análisis de matrices de transición por categorías de riesgo³.

El Cuadro 3 muestra el resultado de las probabilidades de incumplimiento calculadas según el método de matrices de transición por categorías de riesgo deudor.

³ Para efectos del presente documento no se profundizará en el cálculo siguiendo esta metodología, pero para un mayor detalle del cálculo de probabilidades de incumplimiento según matrices de transición se puede consultar (Valencia & Jorge, 2013).

Cuadro 3.
Probabilidades de Incumplimiento por matrices de transición según calificación de riesgo deudor, diciembre 2024

Categorías de Riesgo	Probabilidades de Incumplimiento
A1	1,1%
A2	8,3%
B1	4,4%
B2	39,8%
C1	12,3%
C2	60,1%
D	10,2%
E	94,1%

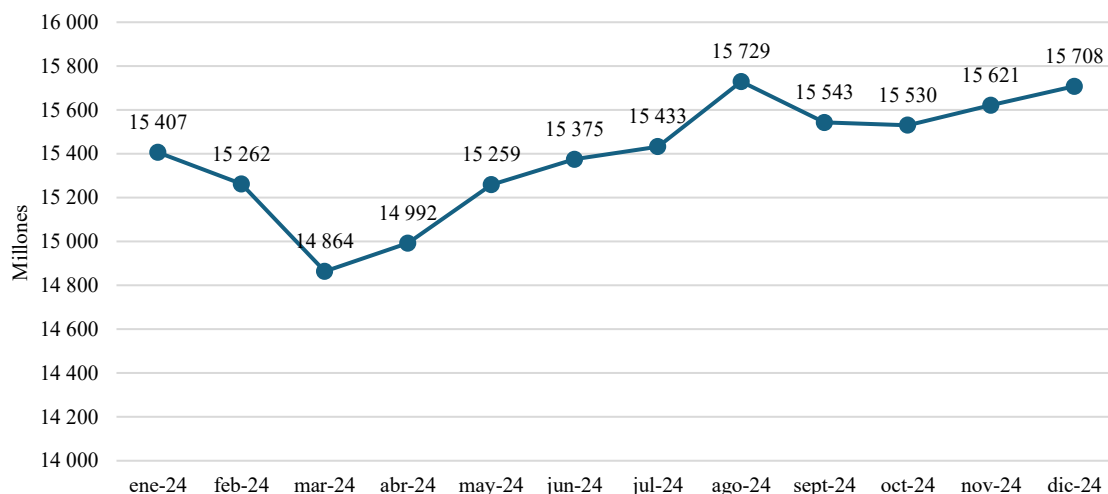
Fuente: JUPEMA

Para este caso, es necesario destacar que uno de los criterios que se buscan para el análisis de las PI, es que aumenten de manera progresiva a medida de que aumenta el nivel de riesgo según categoría, sin embargo el resultado no es así, ya que, si bien la categoría de riesgo A1 (categoría del menor nivel de riesgo), obtiene la menor probabilidad de incumplimiento, y la categoría de riesgo E, obtiene la mayor probabilidad, las categorías intermedias no cumplen con la relación esperada, debido a que, por ejemplo, la categoría de riesgo A2 obtiene una probabilidad de incumplimiento mayor a A1, y a su vez es también mayor al resultado de la categoría B1, cuando este resultado debería ser mayor para la categoría B1, ya que refleja un mayor nivel de riesgo, lo mismo sucede con la categoría de riesgo B2, C2, y D.

El resultado anterior obliga a aplicar ajustes metodológicos para obtener la relación esperada y que pueda ser usado para efectos del cálculo de la PE de la cartera, sin embargo, esto no es deseable, ya que el resultado original en sí sugiere que el indicador es deficiente para medir el constructo que busca explicar, el cual es el nivel de riesgo de las operaciones de crédito.

Ahora bien, una vez calculadas las probabilidades de incumplimiento ajustadas, se llevan al proceso de cálculo de la PE de la cartera. El Grafico 3 muestra el resultado del monto de perdida esperada registrado durante el año 2024, donde se observa que se ha ubicado entre ¢14.800 millones y ¢15.800 millones aproximadamente.

Gráfico 3.
Monto de Perdida Esperada de la cartera de crédito del RCC, enero - diciembre de 2024 (datos en millones de colones)

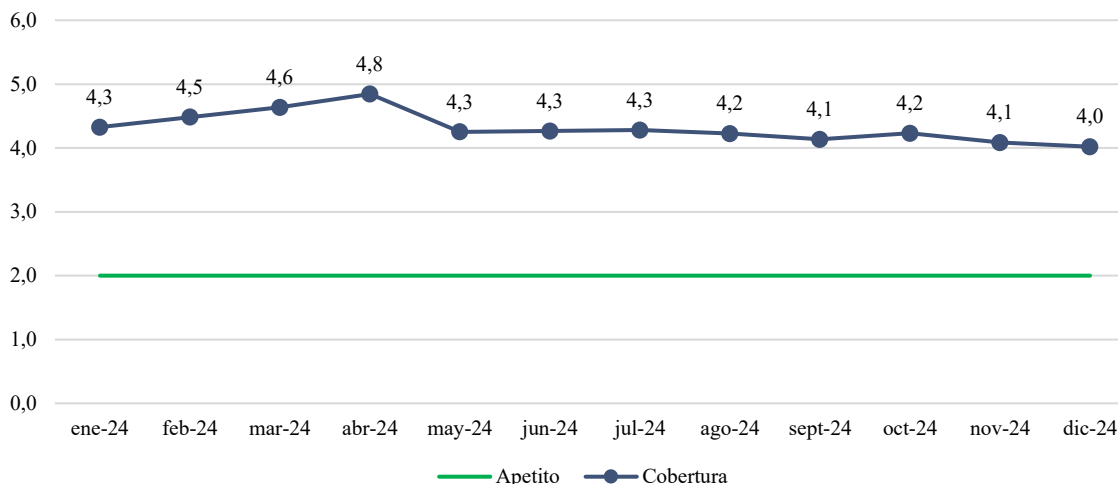


Fuente: JUPEMA

Otro elemento para considerar dentro de la gestión de riesgo de crédito es el nivel de cobertura deseado del monto de perdida esperada. En este sentido, según lo establecido en la Declaración al Apetito al Riesgo institucional, se establecen indicadores por tipo de riesgo, los cuales son comúnmente llamados Indicadores Clave de Riesgo (*KRI* por sus siglas en inglés), los cuales son métricas que sirven para monitorear y medir la exposición a riesgos en una organización, y junto con ellos, se definen límites de apetito y tolerancia al riesgo, con el propósito de llevar un monitoreo de si el nivel de riesgo asumido se encuentra dentro del apetito al riesgo deseado.

Actualmente uno de los KRI definidos dentro de la perspectiva de riesgo de crédito es el nivel de cobertura que otorga el monto de perdida esperada respecto a los saldos de cartera más cercanos a caer en incumplimiento, para este caso, los saldos morosos con más de 90 días de atraso, donde se tiene establecido una cobertura mínima de 2 veces respecto a este saldo. De esta forma, el Grafico 4 muestra el comportamiento del indicador de cobertura KRI durante el año 2024.

Gráfico 4.
Indicador KRI de cobertura de PE respecto a mora con más de 90 días de atraso y límite de apetito al riesgo del RCC, enero - diciembre 2024



Fuente: JUPEMA

Es claro observar, que el resultado de cobertura se ha ubicado dentro del nivel de apetito al riesgo requerido, sin embargo, también salta a la vista que el cumplimiento se ha logrado de manera muy holgada, llegando a alcanzar, más de 4 veces la cobertura respecto a la cartera con mora mayor a 90 días de atraso.

2.5. PROBLEMÁTICA PARA RESOLVER

Según los elementos mostrados anteriormente, la presente investigación pretende abordar dos elementos principales, en primer término, se pretende definir una metodología de calificación de riesgo deudor alternativa que permita establecer probabilidades de incumplimiento que cumplan con los criterios esperados de forma clara y precisa, aportando en una mejor medición del nivel de riesgo de crédito por operación.

En segundo término, existe el cuestionamiento de que la cobertura sobre pérdida esperada que se mantiene actualmente se puede interpretar como una sobre cobertura, en el sentido de que existe la posibilidad de disminuir el monto de pérdida esperada, lo cual podría beneficiar el uso de recursos para posibilidades de inversión, asumiendo un mayor nivel de riesgo, pero con el objetivo de que se mantenga dentro del nivel de apetito definido, por lo tanto también se busca, analizar la viabilidad de que las nuevas probabilidades de incumplimiento puedan

cumplir con este objetivo, y de esta manera establecer si el resultado puede ser aplicado dentro de modelo de gestión de riesgo de crédito institucional.

2.6. REFERENTES O ELEMENTOS TEÓRICOS

En primer término, en lo referente a riesgo de crédito en un contexto financiero, según la definición establecida por el Consejo Nacional de Supervisión del Sistema Financiero (CONASSIF), se entiende como la posibilidad de pérdidas económicas debido al incumplimiento de las condiciones pactadas por parte del deudor, emisor o contraparte, lo que establece la necesidad de realizar una adecuada gestión del riesgo de crédito con el propósito de minimizar las potenciales pérdidas, por lo que según los acuerdos internacionales para la regularización bancaria, Basilea I y II, comúnmente el termino es asociado al concepto de probabilidad de que se materialice el incumplimiento.

Esta relación es incorporada por la Normas Internacionales de Información Financiera número 9 (NIIF 9), cuando establece que las entidades deben aplicar un enfoque prospectivo para evaluar el riesgo de crédito, utilizando las pérdidas crediticias esperadas lo que implica calcular la probabilidad de incumplimiento.

Una de las metodologías más comúnmente utilizadas para definir probabilidades de incumplimiento es la metodología de matrices de transición, tal como menciona Toledo y Leon (2021), el uso de este enfoque tiene diferentes ventajas, como contar con un enfoque sistémico, cuenta con facilidad de cálculo, por lo que no requiere un uso elevado de recursos computacionales y presenta una gran adecuada capacidad predictiva, sin embargo, presenta el inconveniente de que depende de las categorías base para estimar las transiciones, por lo que si dichas categorías no están bien definidas, el resultado podría no ser satisfactorio. Es por esta razón que los autores también mencionan que aparte de las matrices de transición, también se han utilizado diversas herramientas estadísticas para el cálculo de las probabilidades de incumplimiento, como lo son los análisis de modelos de clasificación, modelos de árboles de decisión, análisis de cosechas, entre otros.

En relación con lo anterior, en Costa Rica, los supervisores o reguladores de mercado han incentivado las entidades financieras el desarrollo y exploración de modelos complejos para la gestión de riesgo de crédito, pero se han enfocado más en modelos de originación, los

cuales calculan la probabilidad de incumplimiento de un crédito dadas sus características al momento de otorgamiento para utilizarlo como una herramienta objetiva. A nivel de seguimiento una vez que se ha otorgado el crédito, el desarrollo ha sido más limitado, y se ha mantenido más en modelos de calificación. Actualmente el esquema más reciente vigente está definido según el Acuerdo CONASIF 14-21.

En este sentido, surge el concepto de *score* de crédito, que se enmarca de una manera simple como el resultado de la aplicación de “métodos estadísticos utilizados para clasificar a los solicitantes de crédito, o incluso a quienes ya son clientes de la entidad evaluadora, entre las clases de riesgo bueno y malo” (Hand & Henley, 1997). Por su parte, Escalona (2011) lo describe como el uso de medidas cuantitativas de los resultados de préstamos otorgados en el pasado, para predecir el rendimiento futuro de préstamos con características similares, lo cual ha ofrecido una herramienta de gran importancia en el logro de mayor eficiencia.

A pesar de que en Costa Rica el término de *score* de crédito puede entenderse como de reciente uso, la implementación de estas calificaciones ha sido utilizada desde la década de los años de 1970 (Gutiérrez Girault, 2007), sin embargo, las limitaciones para un adecuado procesamiento de información llevó a que se diera como uso generalizado hasta la década de 1990, cuando el desarrollo de la capacidad computacional permitió un adecuado procesamiento de algoritmos en tiempo real para ser usados como una herramienta en el proceso de otorgamiento de crédito, poco después su avance llevó no solo al uso dentro de la decisión de otorgamiento, sino también al proceso de seguimiento, como una herramienta de monitoreo del nivel de riesgo de la operación durante el ciclo de vida del crédito.

Más concretamente, un *score* de crédito otorga una medida, calificación o escala que permite distinguir entre buenos y malos deudores, o bien se busca tener una estimación de la probabilidad estadística de que el deudor pueda incumplir con sus obligaciones crediticias, lo que se conoce como probabilidad de default (PD) o probabilidad de incumplimiento (PI), por lo que, una probabilidad cercana a uno indica una alta PI y por ende un alto riesgo de crédito, y una probabilidad cercana a cero indica una baja PI y un bajo riesgo de crédito.

Ahora bien, a pesar de que el uso de un *score* de crédito se ha llevado a la práctica de forma amplia en el momento de otorgamiento, resulta razonable pensar que este mismo enfoque

puede ser utilizado para efectos de seguimiento, según la evolución de las características del crédito luego de ser colocado.

A pesar de que la normativa dada por el Acuerdo CONASSIF 14-21 permite el uso de modelos internos condicionados, que se podrían estar aplicando, para efectos de la presente investigación no fue posible ubicar referencias del uso de este enfoque en Costa Rica en otras entidades, lo que se podría explicar debido a que las entidades de carácter financiero son obligadas a seguir el esquema de seguimiento definido por CONASSIF, la cual establece en primera instancia, una metodología basada en indicadores de clasificación, sin embargo, como se muestra más adelante, existen referencias en otros países, enfocadas en ejercicios académicos que han arrojado buenos resultados, lo que abre la posibilidad al uso de este enfoque a futuro.

Existe una amplia gama de literatura que presenta el uso de diversas técnicas o metodologías estadísticas que se han empleado, empezando con modelos considerados más básicos, como análisis de componentes principales y análisis discriminante, pasando por modelos logit y probit, hasta métodos más sofisticados de inteligencia artificial y redes neuronales.

En el trabajo realizado por Granja y García (2020), se busca un modelo de seguimiento de riesgo de crédito para una cartera de clientes de una entidad financiera de en Colombia, para ellos, se utilizan técnicas de minería de datos como el análisis de clúster y árboles de decisión. Dentro de los resultados, se determina que el análisis del modelo permite a la entidad confiar en la estrategia de seguimiento, ya que este clasifica correctamente a los clientes buenos y malos, lo que permite contar con alertas tempranas según su impacto y plantear planes de acción para disminuir los incumplimientos en la cartera.

Por su parte, Cahueñas (2024), lleva a cabo un estudio para desarrollar un *score* de seguimiento para una Cooperativa de Ahorro y Crédito, explorando la metodología de redes neuronales, obteniendo que el modelo se adapta bien a los datos utilizados, logrando un nivel de clasificación correcta de 83%, por lo que se concluye que se puede reconocer la capacidad de modelos de este tipo para contribuir significativamente a mejorar el rendimiento de la cooperativa de estudio.

En el trabajo de García (2023), se hace uso de una gama de algoritmos más amplia para la evaluación de riesgo crediticio para pequeñas y medianas empresas. En la investigación se utilizan métodos de regresión logística, máquinas de soporte vectorial (SVM por sus siglas en inglés, *support vector machine*), “K” vecinos más cercanos, bosques aleatorios, y modelos de *gradient boosting*. Dentro de los principales resultados, se obtiene que el modelo de máquinas de soporte vectorial arroja el mejor balance según los indicadores clasificación y se determina una mejora en la disminución de transacciones en default asociado al modelo utilizado, por lo que se considera oportuno profundizar en el uso de modelos de aprendizaje supervisado.

Se tiene el caso de Vargas y Mostajo (2014), donde mencionan el uso de tres metodologías para la estimación de un *score* crediticio, utilizando modelos de regresión lineales multivariados, modelos *logit* y redes neuronales artificiales.

En su estudio Andreeva (2006), ejemplifica el uso de técnicas demográficas en el modelado de la probabilidad de default por medio de un enfoque de análisis de sobrevivencia, mientras que el uso de árboles de decisión ha sido ampliamente utilizado desde los inicios de los modelos de *score* de crédito, tal como lo muestra el artículo de Venkat y Kim (1987).

Uno de los trabajos que se enfocan en determinar qué clase de modelos pueden ser más eficientes es el de Angarita y Marin (2019), donde se busca realizar una comparación entre modelos tradicionales y de aprendizaje de máquinas (*machine learning*) para el cálculo de probabilidades de incumplimiento, y se determina que los modelos avanzados permiten mejores resultados en materia predictiva.

De manera similar, Cortes (2018) examina los resultados de la comparación entre modelo de regresión logística y modelaje no paramétrico según SVM en un contexto financiero, concluyendo que, aunque el uso de los modelos específicos siempre va a depender del analista, los modelos no paramétricos pueden ofrecer una mejor solución para este tipo de problemas, destacando la ventaja de que no es necesario realizar supuestos de normalidad.

Un estudio más amplio, y en la línea de un esquema de minería de datos similar al propuesto en el presente documento, es desarrollado por Baesens (2003) donde realiza un estudio

aplicando hasta 17 modelos distintos, con 8 bases de datos, con el propósito de evaluar la precisión de los diferentes métodos y determinar los mejores modelos a utilizar, utiliza modelos de regresión logística y programación lineal, diferentes versiones de máquinas de soporte vectorial, y árboles de decisión, dos variantes del método de “K” vecinos más cercanos, redes neuronales y técnicas bayesianas.

Adicional a lo anterior, Peña (2013) recalca que a diferencia de ejercicios de scoring de otorgamiento, en enfoques de seguimiento, para estimar la probabilidad de incumplimiento se debe analizar el comportamiento del deudor hasta que la institución lo considere como pérdida total.

Dentro de los beneficios que pueden otorgar el uso de modelos multivariados se encuentra la posibilidad de definir de manera específica, las probabilidades de incumplimiento de las operaciones de crédito. Según menciona Tamara (2016), las entidades financieras han venido desarrollando sus modelos de originación y seguimiento, que además de ser herramientas para la medición del riesgo de crédito, se han convertido en un insumo de primer orden para definir su estrategia de negocios. En su trabajo, plantea desarrollar un modelo de análisis discriminante para clasificar deudores según su probabilidad de default en una entidad financiera. Dentro de los principales resultados obtenidos se menciona que se llega a obtener probabilidades de incumplimiento inferiores a las que utiliza la entidad financiera, lo que permite liberar fondos sin comprometer la estabilidad financiera de la institución.

Para todos los autores mencionados se destaca que el uso de las diferentes técnicas dependerá de los objetivos que se persigan, y se resalta que la precisión de un determinado modelo sobre otros depende en gran medida de las bases de datos disponibles y el poder predictivos de las variables, por lo que un modelo puede arrojar pobres resultados en una determinada base de datos, pero ser muy preciso con otra, por lo que explorar una gama de modelos en estos ejercicios es un elemento importante.

Por otra parte, un tema que puede ser importante a considerar en ejercicios de modelos de clasificación en cartera de crédito es el gran desbalance que existe en la variable a predecir, lo cual también se da en otros campos, como la medicina y la biología. Diferentes autores han definido técnicas que pueden dar tratamiento a este tema con el propósito de evaluar si

es posible obtener mejores resultados, entre ellas se pueden mencionar técnicas más conocidas como aplicación de submuestreo o sobre muestreo, o métodos centrados en los modelos, como asignar pesos o penalizaciones mayores a los errores en la clase minoritaria, o manejo de parámetros calibrados para ciertos modelos específicos, entre otras técnicas.

Sobre las técnicas de submuestreo y sobre muestreo, se puede citar el trabajo de Carvalho, Pinho y Bras (2025). En su trabajo resaltan la relevancia de valorar el desequilibrio de clases en problemas de clasificación por el impacto significativo que puede tener este fenómeno en los algoritmos de aprendizaje automático, dicho trabajo realizó una revisión bibliográfica exhaustiva sobre los enfoques recientes.

En este artículo, en el caso de el sobre muestreo, se encuentran métodos donde se busca expandir la clase minoritaria enfocados en casos con mayor dureza de instancia, dado que esta se considera la región más decisiva para el entrenamiento de los modelos, aunque también se mencionan métodos más enfocados en áreas con menor dureza de instancia para no agregar ruidos que puedan afectar el resultado de la clasificación, sin embargo en el caso del submuestreo se presenta la misma dualidad, por lo que la decisión debe tratarse buscando la solución que arroje menos pérdida de información. Además, Carvalho, Pinho y Bras (2025), advierten de que una expansión de la clase minoritaria conduce a una mayor sensibilidad, pero a expensas de la especificidad, lo cual debe valorarse según la aplicación específica, el campo de análisis o las preferencias del analista. Por último, se resaltan dos conclusiones, pese a que estas técnicas pueden aportar para mejorar los resultados de clasificación, equilibrar las clases no mitiga inherentemente todos los sesgos dentro de los datos para variables sensibles como la edad o el género, y se menciona que ningún método garantiza un rendimiento superior en todos los conjuntos de datos probados.

Dentro de los autores que han explorado estos elementos se puede mencionar el trabajo de Araf, Idri y Chairi (2024), donde se realizó una recopilación de la literatura disponible sobre técnicas de aprendizaje sensible a costos para datos médicos desequilibrados, incluyendo dos categorías para realizar una comparación entre métricas de evaluación de modelos tradicionales y métricas de costos, concluyen que se observa una infrautilización de las métricas de costos, lo que resalta la oportunidad de incorporarlas en la investigación,

preferiblemente en combinación de múltiples métricas para la búsqueda de una evaluación exhaustiva del rendimiento de los modelos, especialmente en el contexto médico, donde el costo de clasificación errónea en enfermedades es elevado.

Para el caso de modelos que incluyen parámetros para lidiar con desbalance de datos, en el trabajo de Qing-Yan, Jiang-She, Chun-Xia y Sheng-Cai (2013), parten del uso de modelos de potenciación para resaltar las clases minoritarias y desarrollan una investigación para evaluar la efectividad de utilizar variantes de estos modelos para introducir parámetros de costos para aumentar aún más la potenciación de dichas clases. Realizan una comparación empírica de seis algoritmos de potenciación sensibles a costos, como lo son AdaCost, CSB1, CSB2, AdaC1, AdaC2 y AdaC3, entrenando casi 50 mil modelos de clasificación en 17 conjuntos de datos desequilibrados del mundo real. Los resultados muestran que los algoritmos seriales AdaC generalmente superan a AdaCost y CSB al tratar con conjuntos de datos de diferentes niveles de desequilibrio, pero con una menor efectividad en la medida de que se reduce el desequilibrio, lo que favorece la evidencia de que los modelos de potenciación son una opción favorable para el manejo de datos desbalanceados, especialmente si se introducen parámetros de costos.

Por último, un concepto que es necesario mencionar en cualquier estudio ligado a toma de decisiones basada en modelos matemáticos y estadísticos es el riesgo del modelo. Utilizando como referencia la guía de supervisión sobre la gestión del riesgo de modelo publicada por la Reserva Federal y la Oficina del Interventor de la Moneda (OCC, por sus siglas en inglés) de Estados Unidos en el año 2011, el riesgo del modelo se entiende como un mal funcionamiento de un modelo para la toma de decisiones lo cual puede derivar en pérdidas financieras por una mala asignación de recursos.

Según la guía, el riesgo del modelo se puede dar por una serie de causas, entre ellas se pueden mencionar desde un mal diseño metodológico para abordar el problema a resolver, datos inconsistentes o no confiables, mala definición de supuestos al ser erróneos o poco realistas, mala implementación para atender las necesidades al utilizarse en condiciones diferentes respecto al escenario como que fue calibrado, hasta mala interpretación de los resultados. De esta manera, este riesgo debe ser reconocido y gestionado con la misma importancia con que

se gestiona cualquier otro tipo de riesgo, lo que implica identificar causas y generar controles que mitiguen su probabilidad de ocurrencia o su impacto.

Según lo señalado en la guía, un marco de validación eficaz para mitigar el riesgo del modelo debe incluir tres elementos fundamentales:

- Evaluación de la solidez de los conceptos a tratar.
- Monitoreo continuo de los resultados, incluyendo verificación del proceso de análisis y evaluación comparativa.
- Análisis de resultados.

Lo anterior se puede lograr implementando revisiones o auditorías periódicas de los modelos, lo que incluye el monitoreo constante de los resultados según el objetivo esperado, el uso de pruebas retrospectivas, análisis de sensibilidad, análisis de estrés y comparación con modelos alternativos, lo que permite identificar desviaciones o sesgos significativos que puedan ser corregidos, además de generar normas o procedimientos para el registro y manejo de datos para garantizar su confiabilidad y calidad, así como lineamientos base para el abordaje de problemas mediante desarrollo e implementación de modelos.

CAPÍTULO III. METODOLOGÍA

3.1. FUENTE DE DATOS

JUPEMA cuenta con procesos de actualización de información automatizados para el monitoreo y seguimiento de variables relevantes, y se tiene acceso a información desde enero de 2016 a diciembre 2024, de esta manera, las bases de datos a utilizar consisten en información que describe la situación socioeconómica del deudor dueño de la operación, así como información relacionada a sus obligaciones crediticias.

La información socioeconómica se recolecta al momento de otorgar el crédito al deudor y es posible capturar los cambios que se puedan dar en el tiempo, mientras que las características sobre sus obligaciones crediticias se registran de forma automática, tanto para los niveles de endeudamiento, como también su comportamiento de morosidad en la medida de que las operaciones maduren en su ciclo de vida.

Los procesos de actualización de información se implementan mensualmente, por lo que se cuenta con una base de datos de operaciones de crédito donde se puede dar seguimiento a los cambios sufridos en sus características de manera mensual hasta el momento en que dichas operaciones son canceladas, ya sea de forma prematura, si se extinguen dentro del proceso normal de pago, o bien si caen en incumplimiento al ser reconocidas como incobrables.

De esta manera la base de datos está conformada en sus filas por las observaciones de estudio que representan operaciones crediticias, y en sus columnas se tienen las diferentes variables de análisis, como condición particular, salvo excepciones relevantes se busca idealmente tomar como variables aquellas características que puedan cambiar en el tiempo. Además, al ser una base de datos de seguimiento, se define que las bases de datos cuentan con observaciones repetidas, ya que se tiene acceso a las características de una misma operación crediticia en distintos momentos del tiempo.

Es importante mencionar que en problemas de este tipo generalmente se tiene muchas buenas operaciones y pocas malas, lo que genera que la variable de estudio sea desbalanceada (muchas buenas operaciones y pocas malas operaciones), se definirá dentro del proceso de análisis si se utiliza la totalidad de las observaciones disponibles, o una muestra de ella según lo más conveniente para obtener mejores resultados.

Por otra parte, inicialmente se tomará la totalidad de las variables disponibles para ser sometidas a un proceso de análisis de poder predictivo, para tomar en el proceso de modelado únicamente aquellas variables con suficiente poder de discriminaciones entre buenas y malas operaciones de crédito.

Por último, con el objetivo de proteger la confidencialidad de los datos todas las variables identificadoras como números de cedula, segmento de mercado, o números de operación de crédito serán utilizados de manera enmascarada.

3.2. DEFINICIÓN DE VARIABLES DE ESTUDIO

Dentro de las distintas variables a ser consideradas en el estudio según tipo son las siguientes:

VARIABLES DE IDENTIFICACIÓN:

- **Número de operación:** es el número de la operación que se le asigna al crédito solicitado por el afiliado, para efectos de confidencialidad de la información esta variable esta enmascarada.
- **Cedula del deudor:** se refiere al número de cedula del deudor dueño de la operación crediticia, para efectos de confidencialidad de la información esta variable esta enmascarada.
- **Fecha constitución:** es la fecha otorgamiento del crédito al deudor.

VARIABLES CUALITATIVAS:

- **Segmento de mercado:** en esta variable se registra el tipo de institución en que labora el deudor dueño de la operación y su condición laboral, dentro de las instituciones que son cotizantes al RCC se tienen a afiliados pensionados, trabajadores del Ministerio de Educación Pública, trabajadores de Universidades Públicas, del Instituto Nacional de Aprendizaje y de Centros Educativos Privados, para efectos de confidencialidad de la información esta variable esta enmascarada, salvo para el caso del segmento de pensionados.
- **Tipo de Crédito:** Se refiere a si el crédito obedece a un crédito de consumo o uno enfocado a solución de vivienda. Para el caso de créditos de consumo puede incluir

efectos de refundición de deudas, mientras que para créditos de vivienda obedece a planes de compra de lote, construcción, remodelación, o compra de casa.

- **Garantía:** se registran el tipo de garantía que posee los créditos personales solicitados por los afiliados, los cuales pueden tener garantía fiduciaria, hipotecaria o garantía no real (capacidad de pago).
- **Tenencia de hipotecas:** esta variable dicotómica indica si el deudor posee uno o más créditos con garantía hipotecaria, ya sean créditos de consumo o hipotecarios.
- **Tenencia de incobrables:** es una variable dicotómica que indica si el deudor ha mantenido créditos considerados incobrables en el pasado.
- **Plan de inversión:** se registra el objetivo de crédito del deudor, este se divide en las siguientes categorías, “personal” cuando el 100% del monto desembolsado no se utiliza para refundición de deudas, “externo” cuando se utiliza una fracción del monto para refundición de deudas a otras instituciones financieras, pero no internas, “interna” cuando se utiliza una fracción del monto para refundición de deudas internas a JUPEMA, pero no externas, “mixto” cuando se utiliza para la refundición de créditos internos y externos, y vivienda cuando el propósito del crédito es buscar una solución de vivienda.

Variables cuantitativas:

- **Saldo de la operación:** es el saldo total adeudado por la operación crediticia al momento de la medición.
- **Saldo deudor:** es la sumatoria de todas las deudas que tiene el deudor. Se calcula a partir del saldo de cada operación de crédito.
- **Monto cuota:** es el monto de la cuota que debe pagar el deudor por el crédito solicitado al momento de la observación.
- **Suma de cuotas:** es la sumatoria de todas las cuotas de las operaciones de crédito que tiene el deudor.
- **Diferencia de cuotas:** es la diferencia entre la cuota al momento de la medición y la cuota original del crédito al momento de otorgamiento.

- **Valor tasa de interés:** es la tasa de interés vigente del crédito.
- **Cuotas pendientes:** es el total de cuotas que le restan al deudor para la cancelación del crédito al momento de la observación.
- **Amortización:** es el porcentaje del capital prestado que ha sido cancelado por el deudor.
- **Antigüedad:** es la diferencia entre la fecha de otorgamiento del crédito y la fecha de medición. Se expresa en años.
- **Días de mora:** es la cantidad de días de mora que mantiene la operación de crédito al cierre de cada mes.
- **Mora promedio:** es la mora promedio del deudor dueño de la operación crediticia al cierre de mes respecto al total de operaciones que mantiene activas en cartera. Se mide en días.
- **Puntaje deudor:** es un indicador calculado según el Acuerdo CONASSF 14-21 basado únicamente en operaciones de crédito internas a JUPEMA. Se refiere a un indicador que varía de 1 a 5 donde entre más alto indica mayor nivel de morosidad histórica, se calcula según una ventana histórica de hasta 4 años.
- **Mora máxima:** es la mora máxima del deudor con la entidad en los últimos 2 años, considerando el total de operaciones mantenidas por el deudor dueño de la operación.
- **Salario bruto:** es el salario bruto del deudor sin aplicar ningún tipo de deducción.
- **Salario neto:** es la diferencia entre el salario bruto y las deducciones sociales correspondientes según el caso.
- **Salario líquido:** es la diferencia entre el salario neto y la suma de las cuotas correspondientes a todos los créditos que mantenga en JUPEMA.
- **Capacidad de pago:** es el cociente entre el salario líquido y el salario bruto.
- **Cantidad de operaciones:** es la cantidad de operaciones activas que mantiene el deudor con la entidad al momento de la medición.

Variable para predecir:

- **Pagador:** en esta variable se registra si la operación ha caído en incumplimiento al ser reconocida como incobrable en un periodo de 1 año en el futuro respecto al momento de la medición, en caso de que la operación no se haya reconocido como incobrable se reconoce como perteneciente a un buen pagador, en caso contrario se cataloga como perteneciente a un mal pagador⁴.

Según lo anterior, en total se tienen 28 variables iniciales. Además, se descarta utilizar variables demográficas, como el género del deudor, edad, educación y estado civil, debido a las implicaciones negativas de imagen que podría causar en la población afiliada.

3.3. EVIDENCIAS DE CALIDAD DE LA MEDICIÓN PARA LAS VARIABLES DEL ESTUDIO

Tal como se mencionó anteriormente, los datos a utilizar proceden de los registros mantenidos por JUPEMA para su cartera de crédito, para los casos de las variables de identificación, crediticias, endeudamiento, comportamiento de pago y de la variable a predecir, se tienen disponibles procesos automatizados que se ejecutan de forma mensual, por lo que se tiene garantía de la calidad y confianza de los datos.

Para el caso de las variables socioeconómicas, el proceso de actualización depende de la captura de dichas características al momento de realizar el pago de las cuotas crediticias, donde dicho proceso se realiza según deducción de planilla, sin embargo se pueden dar casos donde se den rezagos de actualización cuando se den cambios de condición laboral, centro de trabajo o nivel salarial, pese a lo anterior se consideran datos confiables en el sentido de que dichos rezagos no se mantienen por periodos prolongados debido al proceso de Gestión de Cobro, que considera la actualización de las variables de interés.

Una limitación de información a considerar es que para los casos de variables de endeudamiento y de comportamiento de pago, únicamente se tienen registros de seguimiento

⁴ Por política interna, el Departamento de Crédito y Cobro determina si una operación es reconocida como incobrable tras su análisis individual respectivo, sin embargo, en la gran mayoría de los casos para que una operación de crédito pueda ser reconocida como incobrable debe cumplir con al menos una de las siguientes condiciones: tener más de 120 días de atraso, no recibir pago de cuotas en al menos 3 meses consecutivos, el deudor ha sido declarado con insolvencia económica ante el juzgado judicial, la operación ha sido pasada a cobro judicial.

de las obligaciones crediticias internas con la cartera de crédito del RCC, no así con obligaciones crediticias en el sistema financiero nacional, u otras obligaciones como pagos por pensión alimenticia, lo cual establece una oportunidad de mejora importante para la institución a futuro.

3.4. MODELOS ESTADÍSTICOS DE ANÁLISIS POR EMPLEAR

En esta sección se presentan los principales modelos a ser utilizados en la investigación, los cuales se basan en las referencias bibliográficas mostradas anteriormente, y se enfocan en modelos de aprendizaje de máquinas.

Según menciona García (2023), el aprendizaje de máquinas se presenta como parte de la familia de modelos generados según inteligencia artificial, cuyo objetivo es desarrollar algoritmos que permitan aprender a partir de los datos, por lo que es deseable contar con el mayor volumen de información para el proceso de aprendizaje y de esta manera obtener el mejor resultado.

El proceso de aprendizaje de máquinas busca identificar patrones o relaciones inherentes en las variables de análisis haciendo uso de diferentes técnicas estadísticas y de esta forma poder realizar predicciones o pronósticos utilizando datos nuevos, los cuales no fueron utilizados para el proceso de aprendizaje. La diferencia con los enfoques clásicos frecuentistas radica en que no es necesario suponer de previo que existe un proceso dado que genera los datos para llegar a una solución. Existen dos grupos de modelos de este tipo, los modelos de aprendizaje no supervisado y aprendizaje supervisado.

3.4.1. Aprendizaje no supervisado

En los modelos de aprendizaje no supervisado se busca aplicar un algoritmo a una base de datos, sin definir previamente etiquetas o categorías en las variables, con el propósito de obtener patrones en los datos según una distribución de probabilidad subyacente, los cuales suelen no ser fáciles de detectar. Estas técnicas suelen utilizarse para distintos propósitos, como lo pueden ser conformación de grupos de observaciones (clúster) según la mayor similitud entre ellos, reducción de dimensiones cuando se tienen muchas variables, pero manteniendo la mayor cantidad de información posible, entre otras técnicas.

En este sentido, para la presente investigación se muestran dos técnicas de interés que tienen el propósito de servir para explorar la relación de las variables, su capacidad predictiva y la posibilidad de servir para crear indicadores que sirvan como nuevas variables predictoras en los modelos de clasificación, los cuales son los siguientes modelos:

3.4.1.1. Análisis de Componentes Principales

Tal como menciona Hernández (2013), el propósito del análisis de componentes principales (ACP) es reemplazar p variables X_j ($j = 1, 2, \dots, p$), generalmente correlacionadas, por un número mucho menor de variables no correlacionadas, llamadas componentes principales, de forma que se conserve la mayor parte de la varianza posible, las nuevas variables (componentes) estén no correlacionadas y los vectores de proyección sean ortogonales, para de esta forma explicar la mayor porción de la variabilidad total de las p variables.

Los componentes principales se construyen como combinaciones lineales particulares de las X_j de tal forma que además de no estar correlacionados, tengan también variancia máxima, para de esta manera simplificar el análisis y la interpretación.

El objetivo del modelo es encontrar un vector w_1 tal que la proyección $z_i = w_1^T x_i$ tenga variancia máxima bajo la restricción $\|w_1\| = 1$. La variancia está dada por:

$$Var(z) = \frac{1}{n} \sum_{i=1}^n (w_1^T x_i)^2 = w_1^T S w_1 \quad [2]$$

donde $S = \frac{1}{n} X^T X$ es la matriz de covariancia muestral. El componente w_1 se obtiene por medio de la solución por multiplicadores de Lagrange, mientras que los siguientes componentes siguen el mismo proceso bajo la condición de que sean ortogonales. De esta manera se obtiene un vector W_k que permite realizar una proyección en un espacio reducido en dimensiones.

Por último, la variancia total del conjunto de datos se puede expresar como:

$$Var \text{ total} = tr(S) = \sum_{i=1}^d \lambda_i \quad [3]$$

Por lo que la proporción de la varianza explicada por el componente k estaría dada por:

$$Var\ explicada = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^d \lambda_i} \quad [4]$$

Lo anterior se utiliza para decidir cuantos componentes es necesario utilizar.

Cabe mencionar que el ACP puede ser también muy útil como una primera etapa de un análisis más completo, en el que técnicas multivariadas que utilicen como variables de entrada a los componentes principales, también sirve para detectar la presencia de observaciones multivariadas atípicas, entro otros usos.

3.4.1.2. Método de conformación de grupos K-medias

De la misma manera, Hernández (2013) menciona que el análisis de agrupamientos (“*clúster analysis*”) es una técnica estadística que sirve para clasificar individuos u objetos en grupos relativamente homogéneos.

La técnica se basa en una medición de la “proximidad” o semejanza entre los individuos u objetos, donde hay muchas opciones para realizar esta medición, y en las aplicaciones de la técnica el investigador puede escoger la medida de proximidad o semejanza que se considere apropiada. Para este caso se utiliza la técnica de K medias, por el motivo de que una técnica de agrupamiento no jerárquico suele dar buenos resultados cuando se tiene una gran cantidad de observaciones.

Para este método se parte de especificar el número de grupos “K”, lo que define los “K” grupos iniciales. Cada grupo tiene como centro o centroide⁵ el vector de p observaciones correspondiente a las p variables del caso que define al grupo, y se procede luego a asignar cada uno de los restantes n-K casos, uno a la vez, al grupo cuyo centro esté más cerca de él, esto se realiza según un proceso iterativo mediante el cual, el proceso anterior se repite hasta determinar los centros finales de los K grupos. El proceso acaba cuando ya no se producen más cambios en los centros de los grupos, o cuando se ha alcanzado un número predeterminado de iteraciones.

⁵ El centroide es el punto que representa el centro de cada grupo, y se calcula como la media de todos los puntos de datos asignados a ese grupo.

Su formulación matemática se puede describir como:

Dado un conjunto de datos $X = \{x_1, \dots, x_n\}$ se busca dividir las observaciones en “K” grupo C_1, \dots, C_k de forma que se minimice la suma del error cuadrático dentro de cada clúster, donde se define una función de costo “J” tal que:

$$J = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad [5]$$

Donde μ_k se refiere al centroide del grupo C_k .

Por último, el objetivo es encontrar una asignación C_1, \dots, C_k y los centroides μ_1, \dots, μ_k de modo que minimice la función de costo “J”.

3.4.2. Aprendizaje supervisado

Tal como comenta García (2023), a diferencia del caso anterior, los modelos de aprendizaje supervisado utilizan variables ya clasificadas para la construcción de los modelos, las cuales son llamadas variables respuesta, esto debido a que se busca obtener la mejor estimación posible de la función $f(x)$ para todo x , donde x se compone del vector de variables explicativas, y que dicha función se pueda aplicar a un nuevo conjunto de datos para predecir la variable respuesta no observada.

En la estimación de los modelos de este tipo se busca minimizar la función de pérdida empírica, lo que significa minimizar en los datos de aprendizaje, la diferencia entre las predicciones que arroja el modelo y los valores reales observados.

Dependiendo de la naturaleza de la variable respuesta se pueden calcular indicadores de eficiencia distintos, si se tiene un caso de una variable continua, se pueden calcular medidas como el error cuadrático medio, mientras que en casos donde la variable respuesta es categórica, se puede medir según la proporción de observaciones clasificadas correctamente.

Para el caso analizado en el presente documento, para establecer si los resultados de los modelos resultan adecuados, se hace uso de medidas de clasificación correcta, debido que la variable respuesta está representada como una variable categórica dicotómica, el detalle de estas medidas se muestra más adelante en la definición del plan de pruebas.

A continuación, se muestran los modelos predictivos de clasificación que se consideran más apropiados para cumplir con el objetivo del estudio, donde se valoró el alcance y el tiempo disponible.

3.4.2.1. K vecinos más cercanos

Según lo mostrado en el desarrollo de Hastie, Tibshirani, y Friedman (2008), es un algoritmo de clasificación que tiene como objetivo identificar los K vecinos más cercanos de un punto de consulta dado según los datos de aprendizaje, de modo que podamos clasificar ese punto.

En el caso de clasificación binaria, para los K vecinos más cercanos en la muestra de aprendizaje o entrenamiento, el predictor y_0 estima la probabilidad condicional para la primera clase como la fracción de observaciones en K para la cual $y_j = 1$, según la siguiente manera:

$$\widehat{y}_0 = \begin{cases} 1 & \text{si } \widehat{P}(y_0 = 1 | x = x_0) = \frac{1}{K} \sum_{j \in K} I(y_j = 1) > h \\ 0 & \text{si } \widehat{P}(y_0 = 1 | x = x_0) = \frac{1}{K} \sum_{j \in K} I(y_j = 1) < h \end{cases} \quad [6]$$

En general, la clasificación de un vector de variables explicativas x_i depende del predictor y_i , que viene dado por la siguiente expresión:

$$\widehat{y}_0 = \begin{cases} 1 & \text{si } \frac{1}{K} \sum_{j \in K} y_j > h \\ 0 & \text{si } \frac{1}{K} \sum_{j \in K} y_j < h \end{cases} \quad [7]$$

Siendo h el umbral de clasificación definido, generalmente tomando un valor de 0,5.

Donde la observación j es vecino de i si la distancia entre los vectores de características x_i y x_j está entre las menores de toda la muestra de entrenamiento.

Existen varias posibilidades para el cálculo de las distancias⁶, para el presente documento se considera la distancia más utilizada, la cual es la distancia Euclidiana, la cual se expresa de la siguiente forma:

$$\text{Distancia euclidiana} = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad [8]$$

3.4.2.2. Regresión logística

Tal como lo presenta Hastie, Tibshirani, y Friedman (2008), la regresión logística funciona para ejercicios de aprendizaje supervisado cuando la variable respuesta es dicotómica, lo cual permite modelar la probabilidad de ocurrencia del evento entre las dos clases posibles. La formulación matemática está dada por la función logística, que parte del modelo clásico de regresión lineal, para modelar la relación entre la variable respuesta binaria “y”, con las variables explicativas “x”, con el propósito de estimar la probabilidad de observar $y=1$ dado x . Por lo tanto:

$$p(x) = P(y = 1|x) = x\beta \quad [9]$$

Donde x es el vector de variables explicativas y β es un vector de k componentes. Luego, con el propósito de delimitar el resultado para que tome valores entre cero y uno se aplica la función logística al modelo lineal tal que:

$$P(Y = 1|X_K) = \frac{e^{x\beta}}{1 + e^{x\beta}} = \frac{e^{(a+b_1x_1+\dots+b_kx_k)}}{1 + e^{(a+b_1x_1+\dots+b_kx_k)}} \quad [10]$$

Posteriormente, aplicando logaritmos a ambos lados de la igualdad se obtiene la siguiente expresión:

$$\log\left(\frac{p(x)}{1 + p(x)}\right) = x\beta \quad [11]$$

Por último, la función predictora para la i -ésima observación viene dada por:

⁶ Existe variedad de medidas de distancia que pueden ser utilizadas, entre ellas se pueden mencionar, Minkowski, Manhattan, Cosine, Jaccard, entre otras. Para una mayor referencia se puede consultar Arfat (2019)

$$\hat{y}_i = f(x_i) = \begin{cases} 1 & \text{si } \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} > h \\ 0 & \text{si } \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \leq h \end{cases} \quad [12]$$

Siendo h el umbral de clasificación definido, generalmente tomando un valor de 0,5. Además, los componentes del vector β se estiman mediante la maximización de la función de verosimilitud.

3.4.2.3. Máquinas de soporte vectorial (SVM)

Tal como se muestra en Hastie, Tibshirani, y Friedman (2008), al igual que el caso anterior, el algoritmo SVM se utiliza para la clasificación y regresión de datos dicotómicos. Este método se basa en tratar de encontrar el mejor hiperplano que separe a las categorías de una variable de interés con el mayor “margen” posible para hacer predicciones precisas.

Para este algoritmo las observaciones clasificadas como 0 en la base de datos original se reclasifican como -1 para su funcionamiento, de este modo, el hiperplano que separa las clases entre $y = 1$ o $y = -1$ se define como:

$$x\delta = 0 \quad [13]$$

Donde δ es un vector normal al hiperplano, y la ecuación anterior representa una ecuación general del mismo. Posteriormente, la regla de clasificación para la observación i dado su vector de características x_i estaría dado por:

$$\hat{y}_i = \begin{cases} 1 & \text{si } x_i\delta \geq 1 \\ -1 & \text{si } x_i\delta < 1 \end{cases} \quad [14]$$

Por otro lado, la distancia entre el hiperplano y los puntos de datos más cercanos se conoce como margen, donde el algoritmo SVM busca maximizar el margen para encontrar la mejor separación entre las clases, sin embargo, existe la posibilidad de que las clases no sean separables por un hiperplano, debido a la presencia de relaciones no lineales. En este caso, se aplica una transformación de los datos contenidos en el vector de variables explicativas de forma tal que permita obtener una frontera de clasificación lineal. Esta transformación se lleva a cabo mediante la utilización de una función k denominada kernel, para el caso de

SVM se pueden utilizar hasta 4 kernel diferentes, como lo son "linear", "radial", "polynomial", "sigmoid".

3.4.2.4. Método Naive Bayes

Siguiendo lo mostrado por Hastie, Tibshirani, y Friedman (2008), este método se basa en el Teorema de Bayes para definir un algoritmo de clasificación donde se asume que los predictores de un modelo son condicionalmente independientes, lo cual reduce los procesos de cálculo y permite clasificar una probabilidad a cada variable, y se clasifican las observaciones según la probabilidad mayor para cada combinación de variables. Su formulación para un caso dicotómico se resume de la siguiente forma:

A partir de la intención de predecir una clase C_k que maximiza la probabilidad posterior para una observación $x = (x_1, x_2, \dots, x_n)$ donde cada x_i pertenece a $\{0,1\}$ y se tiene:

$$\hat{C} = \arg \max_{C_k} P(C_k|x) \quad [15]$$

Se utiliza el teorema de Bayes donde:

$$P(C_k|x) = \frac{P(x | C_k) * P(C_k)}{P(x)} \quad [16]$$

Considerando que $P(x)$ es constante para todas las clases, se busca maximizar el numerador:

$$\hat{C} = \arg \max_{C_k} P(x|C_k) * P(C_k) \quad [17]$$

Para lo anterior, dado que se tiene independencia lineal entre los diferentes atributos, se usa la máxima verosimilitud de modo que:

$$P(x|C_k) = \prod_{i=1}^n P(x_i|C_k) \quad [18]$$

Para el caso dicotómico, cada atributo x_i es tratado como una variable binaria:

$$P(x_i|C_k) = \begin{cases} \theta_{ik} & \text{si } x_i = 1 \\ 1 - \theta_{ik} & \text{si } x_i = 0 \end{cases} \quad [19]$$

Donde el termino $\theta_{ik} = P(x_i = 1 | C_k)$

Por lo que se obtiene:

$$P(x|C_k) = \prod_{i=1}^n \theta_{ik}^{x_i} (1 - \theta_{ik})^{1-x_i} \quad [20]$$

Y el clasificador se deriva de la siguiente manera:

$$\hat{C} = \arg \max_{C_k} \left[P(C_k) * \prod_{i=1}^n \theta_{ik}^{x_i} (1 - \theta_{ik})^{1-x_i} \right] \quad [21]$$

Por último, para la estimación de las probabilidades a priori de las clases según los datos de aprendizaje se aplica:

$$P(C_k) = \frac{N_k}{N} \quad [22]$$

Donde N_k es el número de eventos de la clase C_k y N es el total de observaciones.

Y se obtiene la probabilidad de cada atributo de la clase tal que:

$$\theta_{ik} = \frac{1 + \sum_{j=1}^{N_k} x_{ij}}{2 + N_k} \quad [23]$$

3.4.2.5. Análisis discriminante lineal (ADL o LDA)

Tal como desarrolla Hastie, Tibshirani, y Friedman (2008), este método se basa en el Teorema de Bayes para clasificar una observación en una determinada clase, con la diferencia de que la probabilidad para asignar la clase será calculada usando la función de densidad de la distribución normal.

Como supuestos del modelo se define se tiene la misma matriz de covarianza para todas las clases, cada clase tiene su propia media y las clases tiene una probabilidad a priori π_k .

De esta manera, si se multiplica la distribución normal multivariada por la probabilidad a priori se obtiene:

$$P(x|C_k) * \pi_k = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)^T \right) * \pi_k \quad [24]$$

Luego, al tomar el logaritmo de la expresión y omitiendo los valores constantes se obtiene:

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + \log \pi_k \quad [25]$$

Desarrollando el cuadrado de la expresión y sustituyendo el valor $\delta_k(x)$ es posible obtener la función discriminante lineal dada por:

$$\delta_k(x) = x^T w_k + w_{k0} \quad [26]$$

Donde: $w_k = \Sigma^{-1} \mu_k$ y $w_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$

Y se asigna x a la clase con mayor discriminante tal que:

$$\hat{C} = \arg \max_k \delta_k(x) \quad [27]$$

3.4.2.6. Árboles de decisión

Tal como desarrolla García (2023), este método es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja, su formulación matemática del algoritmo a utilizar se basa en el método de "árboles de clasificación y regresión" (CART).

Este modelo consiste en la división binaria de forma recursiva de la muestra de aprendizaje en base a una característica determinada, obteniendo distintos nodos internos y un total de l nodos terminales u hojas. En cada una de las hojas del árbol de decisión se obtienen distintas proporciones entre las clases correspondientes a la muestra de entrenamiento, por lo que la función de predicción clasificará a la observación i -ésima que caiga en la hoja k con la clase cuya proporción sea mayor.

Durante la construcción del árbol de decisión se divide el nodo k en dos nodos k_L y k_R en base a una característica observable x_j y un margen t de forma tal que:

$$\begin{aligned} k_L(t, j) &= \{i \mid x_i^j \leq t\} \\ k_R(t, j) &= \{i \mid x_i^j > t\} \end{aligned} \quad [28]$$

El criterio utilizado para la división mide la heterogeneidad o impureza de k_L y k_R a través de la proporción de individuos que pertenezcan a la clase 1 y se encuentren en el nodo k denominada como p_k , por lo que se define lo siguiente:

$$\begin{aligned} D_k^E &= -|k|(p_k \log p_k + (1 - p_k) \log(1 - p_k)) \\ D_k^G &= -2|k|p_k(1 - p_k) \end{aligned} \quad [29]$$

Los términos anteriores reflejan la heterogeneidad medida a través de la Entropía del nodo k y la heterogeneidad medida a través de la concentración de Gini del nodo k respectivamente. De esta forma, las variables j y t se eligen de forma tal que minimicen:

$$D_{kL} + D_{kR} \quad [30]$$

Donde D_{kL} y D_{kR} son las medidas de heterogeneidad del nodo izquierdo y el nodo derecho en la partición a realizar.

Este proceso se lleva a cabo de manera recursiva de nodo a nodo hasta obtener un nodo completamente homogéneo o que se cumpla con un criterio de interés como cantidad mínima de observaciones por nodo, o profundidad del árbol.

Este algoritmo generalmente utiliza la impureza de Gini para identificar el atributo ideal para realizar una clasificación. La impureza de Gini mide la frecuencia con la que se clasifica incorrectamente un atributo elegido al azar.

$$GINI = 1 - \sum [p(j|t)]^2 \quad [31]$$

3.4.2.7. Bosques aleatorios

Siguiendo con el desarrollo de García (2023), en este método se toman “ m ” muestras aleatorias con reemplazo (*Boostraps*) de los datos originales, para luego aplicar en cada una de ellas un método predictivo, en este caso el método de árboles de decisión, y clasificar a la observación en la categoría que obtenga la mayor cantidad de asignaciones en las muestras.

Cabe resaltar que, a diferencia del modelo de árboles de decisión, que puede mostrar alta variabilidad, el algoritmo de Bosques Aleatorios corrige este problema mediante la estimación de un número finito de árboles de decisión, para lo cual considera los siguientes criterios:

- Se obtiene una submuestra por *bootstrapping* con reemplazo de la muestra de entrenamiento.
- Se eligen aleatoriamente m variables explicativas del conjunto de p totales. En general, se selecciona m como la raíz cuadrada de p .
- Se construye un árbol óptimo con la submuestra y las m variables explicativas seleccionadas según lo descrito anteriormente.
- El proceso se repite una cantidad finita de veces.
- Finalmente, el predictor obtenido por el algoritmo de Bosques Aleatorios es el resultado del consenso de mayoría de votos, o, en otras palabras, si en la mayoría de los árboles el predictor toma valor 1, se le asigna ese valor.

3.4.2.8. Métodos de potenciación Ada-boosting

Siguiendo la línea de Hastie, Tibshirani, y Friedman (2008), los modelos de carácter *boosting* presentan un enfoque general que se puede aplicar a distintos métodos de aprendizaje estadístico para clasificación, para este caso, se da un enfoque aplicado según arboles de decisión.

Similar al caso de bosques aleatorios, este método toma una muestra aleatoria de los datos originales y aplica un método clasificador, sin embargo, en este caso se busca aumentar el peso o potenciar a los individuos mal clasificados para que en la siguiente aplicación del método clasificador se enfoque más en estos individuos, mejorando su clasificación. El algoritmo para potenciar los casos mal clasificados se basa en la siguiente relación:

Se definen los pesos en la primera iteración, con igual ponderación de la siguiente manera:

$$w_i = \frac{1}{N}, i = 1, 2, \dots, N \quad [32]$$

Luego se define el clasificador $G_m(x)$ que pertenece al intervalo $(-1,1)$, y se ajusta el clasificador $G_m(x)$ con los datos de entrenamiento utilizando los pesos w definiendo una función de distribución de probabilidad, para esto se calcula el error de los casos clasificados:

$$err_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i} \quad [33]$$

Con lo anterior se calcula el peso para la siguiente iteración para los casos mal clasificados de la siguiente forma:

$$\alpha_m = \log\left(\frac{1 - err_m}{err_m}\right) \quad [34]$$

$$w_i \leftarrow w_i * \exp[\alpha_m * I(y_i \neq G_m(x_i))] , i = 1, 2, \dots, N \quad [35]$$

Y por último se actualiza el clasificador actualizando la función de distribución:

$$G(x) = \text{sign}\left[\sum_{m=1}^M \alpha_m G_m(x)\right] \quad [36]$$

3.4.2.9. Métodos de potenciación XG-boosting

Según Hastie, Tibshirani, y Friedman (2008), Para el año 2000, el estadístico J Friedman, desarrolla una variación del algoritmo *AdaBoost* interpretándolo como procedimientos de estimación por etapas para ajustar un modelo de regresión logística aditiva, demostrando que en realidad el algoritmo *Ada-boosting* estaba minimizando una función de pérdida exponencial, a este algoritmo se le llamó *Gradient-boosting*.

En este enfoque los árboles son creados utilizando información de árboles previos y, a diferencia del algoritmo de Bosques Aleatorios, cada árbol se construye a partir de la muestra de entrenamiento original de la siguiente forma:

- Se estima el mejor árbol de decisión posible para la totalidad de datos de la muestra de entrenamiento, se obtiene la predicción y las observaciones bien y mal clasificadas.
- Se estima el mejor árbol de decisión posible sobre las observaciones mal clasificadas y a su pronóstico se le denomina como f'_1 .
- Se actualiza el predictor para la totalidad de la muestra de la siguiente forma y se computan las nuevas observaciones mal clasificadas:

$$f_1 = \frac{1}{2} f_0 + \frac{1}{2} f'_1 \quad [37]$$

- Se repite el proceso una cantidad m de veces.

La idea detrás del funcionamiento de este algoritmo es que en cada iteración el pronóstico se actualiza para generar predicciones más certeras.

Más adelante, con la contribución del entonces estudiante de la Universidad de Washington, Tianqi Chen (2016), se realizó una mejora creando un sistema escalable del algoritmo *Gradient-boosting*, al que le llamó *XG-boosting*. Chen introdujo el uso de una función objetivo regularizada que se puede expresar de la siguiente manera:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad [38]$$

Con:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad [39]$$

Donde el termino $\Omega(f_k)$ penaliza la cantidad de hojas T y pesos W .

Las mejoras aportadas por Chen permiten varias optimizaciones en su implementación tanto a nivel de su algoritmo como también a nivel su tiempo de entrenamiento, que resulta notablemente más rápido respecto a otros modelos, actualmente es uno de los modelos más usados en problemas de clasificación.

3.5. GENERACIÓN DEL PLAN DE PRUEBAS

Siguiendo un esquema similar al planteado en los trabajos de Etim y Damidele (2022) y García (2023), los conjuntos de datos a utilizar para la estimación de los modelos se dividen de forma aleatoria de la siguiente manera:

- **Base de datos de análisis:** consiste en el total de observaciones y variables con las cuales se procede a realizar los procesos de entrenamiento y prueba de los modelos.
- **Base de datos de entrenamiento:** con esta muestra se estiman los parámetros pertinentes para cada algoritmo a utilizar. Consiste en una selección aleatoria del 80% de la base de datos de análisis.
- **Base de datos de prueba:** con esta muestra se comparan los algoritmos según los indicadores de clasificación definidos y se elige aquel que demuestre mejor capacidad

predictiva. Consiste en el 20% restante de la base de datos de análisis luego de seleccionar la base de datos de entrenamiento.

- **Base de datos de validación:** consiste en una base de datos fuera de los datos de aprendizaje que sirve para la aplicación del mejor algoritmo elegido y la verificación del comportamiento de los pronósticos. En este punto se destaca que evaluar los resultados del modelo en una base de datos fuera de los datos de entrenamiento y prueba busca aportar elementos para verificar que la capacidad predictiva de las variables contenidas en el modelo es efectiva.
- **Base de datos nuevos:** se refiere a una base de datos nuevos con el propósito de estudiar la viabilidad de los resultados obtenidos según los objetivos del estudio.

Como criterios para evaluar la calidad y valides de los modelos a desarrollar se establecen los indicadores de clasificación correcta por categorías, en este sentido al contar con un problema de dos categorías divididas en operaciones pertenecientes a deudores que caen en incumplimiento, o malas operaciones, y operaciones que no lo hacen, o buenas operaciones, las medidas de error se generan a partir de la construcción de una matriz de confusión, las cuales son altamente utilizadas y consisten en una tabla que recoge las clasificaciones que se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas, dicha matriz, así como las medidas de calidad a utilizar están definidas de la siguiente manera:

Cuadro 4.
Representación de matriz de confusión

Operaciones		Pronostico	
		Malas	Buenas
Dato	Malas	a	b
Real	Buenas	c	d

Fuente: Etim y Damidele (2022) y García (2023)

De esta manera, continuando con la referencia de los trabajos de Etim y Damidele (2022) y García (2023), se definen las siguientes medidas de precisión:

- **Porcentaje de precisión global:** se refiere al porcentaje de operaciones clasificadas de forma correcta respecto al total de casos.

$$\text{Precisión Global} = \frac{a + d}{a + b + c + d} \quad [40]$$

- **Porcentaje de clasificación correcta de malas operaciones (incumplimiento) o sensibilidad:** se refiere al porcentaje de casos clasificados de forma correcta que caen en incumplimiento respecto al total de operaciones en incumplimiento.

$$\text{Precisión malas operaciones} = \frac{a}{a + b} \quad [41]$$

- **Porcentaje de clasificación correcta de buenas operaciones (no incumplimiento) o especificidad:** se refiere al porcentaje de casos clasificados de forma correcta que no caen en incumplimiento respecto al total de operaciones que no caen en incumplimiento.

$$\text{Precisión buenas operaciones} = \frac{d}{c + d} \quad [42]$$

- **Curva ROC y área bajo la curva (AUC):** La curva ROC puede entenderse como la representación gráfica de la relación entre la sensibilidad y especificidad, y mide hasta qué punto un método de clasificación es capaz de detectar o clasificar los casos de una clase correctamente (en este caso malas operaciones), de entre todos los casos de esa clase. De la curva ROC en un eje positivo de un plano cartesiano, se permite calcular el indicador de área bajo la curva o (AUC), el cual es un indicador que varía entre cero y uno, donde entre más cercano a uno se obtiene un mejor desempeño de clasificación. Este indicador permite comparar el desempeño entre diferentes métodos de clasificación y permite medir el desempeño según diferentes probabilidades de corte.
- **Curva de Precisión Recuperación y área bajo la curva (AUC-PR):** de manera similar al AUC de la curva ROC, es posible estimar una curva a partir de los indicadores de precisión y recuperación derivados de la matriz de confusión para cada punto de corte. La precisión se entiende como el total de verdaderos positivos entre todas las predicciones positivas, mientras que la recuperación se mide como el total de verdaderos positivos entre todos los positivos reales, lo que equivale a la

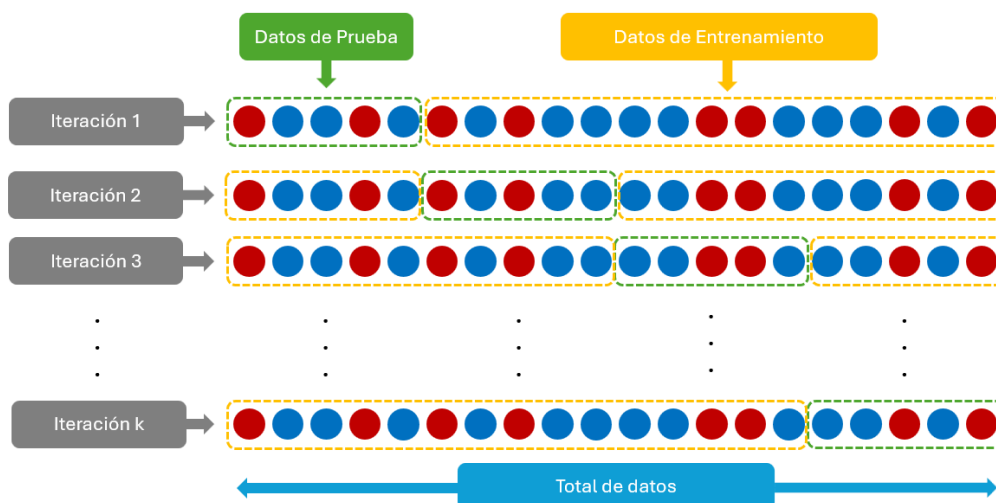
sensibilidad. Si se ubica la precisión en un eje positivo de un plano cartesiano, cuanto más alto sea el eje Y en la curva, mejor será el rendimiento del modelo, también se permite calcular el indicador de área bajo la curva o (AUC-PR), el cual es un indicador que varía entre cero y uno, donde entre más cercano a uno se obtiene un mejor desempeño de clasificación. Este indicador es recomendado cuando se tiene problemas con un gran desbalance de la variable respuesta.

- **Valor F o F1 Score:** es una medida de desempeño de clasificación de un modelo, y se considera como una media armónica que combina los valores de la precisión y de la exhaustividad, es un indicador que varía entre cero y uno, donde entre más cercano a uno se obtiene un mejor desempeño de clasificación.

$$\text{Valor } F = 2 \frac{\text{Precisión} * \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}} = \frac{2a}{2a + b + c} \quad [43]$$

Para la evaluación de los indicadores anteriores para cada modelo, se realiza un proceso de validación cruzada siguiendo un esquema similar al sugerido en Amazon Web Services (2025), y Learn Microsoft (2024), que se resume según lo que se muestra en la Figura 2. Esta técnica consiste en dividir la base de datos en un conjunto de entrenamiento k y al restante $k-1$ como conjunto prueba, repitiendo el proceso k veces o “ k folds”, de modo que cada grupo se utilice una vez como conjunto de prueba para cada una de las combinaciones posibles de parámetros, y cada iteración se realiza el contraste de base de datos de entrenamiento y prueba con un conjunto diferente. De esta manera se garantiza una mayor precisión del error y evita la sobre estimación por factores aleatorios. Para este caso se utilizan 5 iteraciones dividiendo la base de datos en 10 grupos en cada iteración.

Figura 2.
Diagrama de validación cruzada usando K grupos



Fuente: Amazon Web Services (2025)

Cabe resaltar que, siguiendo la misma referencia, para cada modelo empleado se realiza la escogencia de ciertos parámetros de interés según lo mostrado en el Cuadro 5, para determinar los que arrojan los mejores resultados. Se plantea realizar la selección o calibración de dichos parámetros siguiendo el proceso de validación cruzada mencionado anteriormente.

Cuadro 5.
Parámetros evaluados

Modelos	Parámetros
K vecinos	Kernel: "rectangular", "triangular", "epanechnikov", "biweight", "triweight", "cos", "inv", "gaussian", "rank", "optimal"
SVM	Kernel: "linear", "radial", "polynomial", "sigmoid"
Ada bosting	Booster: "discrete", "real", "gentle"
XG boosting	Booster: "gbtree", "gblinear", "dart"
Arboles de decisión	Mínimo de cada nodo
Bosques aleatorios	Cantidad de árboles

Fuente: elaboración propia a partir de Datos de Crédito de JUPEMA

De esta manera, como criterios de aceptación de un modelo adecuado se establece lo siguiente:

- Identificar con el mejor modelo obtenido al menos un 60% de las operaciones crediticias que pertenecen a malos pagadores o caen en incumplimiento.
- Alcanzar con el mejor modelo obtenido un porcentaje de precisión de buenas operaciones crediticias y de precisión global de al menos 85%.

Otro elemento para considerar es que, por defecto, la clasificación entre buenas y malas operaciones se realiza con una probabilidad de 0,5 (probabilidad de corte 0,5), sin embargo, adicionalmente se realizará un análisis de la mejor probabilidad de corte de clasificación entre buenas y malas operaciones, con el propósito de obtener los mejores resultados posibles para el mejor modelo seleccionado.

Por último, como paso final en el proceso del plan de pruebas y validación, una vez que se tiene el modelo que arroja los mejores resultados, se toma una base de datos de casos nuevos, ajenos a la base de datos de entrenamiento y prueba, y se evalúan los pronósticos según el modelo aplicado en función de que las probabilidades de incumplimiento generadas cumplan con los criterios esperados.

CAPITULO IV. RESULTADOS

4.1. ANÁLISIS DESCRIPTIVO

En este apartado se presentan los resultados del análisis realizado. Inicialmente, se presentan las principales estadísticas descriptivas del estudio. Se cuenta con un total de 67.813 operaciones de crédito correspondientes a créditos consumo y vivienda, colocadas a principios y a finales del año 2023 (enero y diciembre), y se dividen según las variables explicativas entre buenas y malas operaciones, esta base de datos conforma los datos de entrenamiento y prueba. Para efectos de selección de las observaciones que conforman la base datos de entrenamiento, se seleccionan las mediciones de inicio y final del año 2023, con el propósito de manejar una base de datos que refleje en la mayor medida posible, los cambios que se pueden dar en las variables explicativas.

El Cuadro 6 muestra la distribución de las operaciones según las variables cualitativas, junto con el porcentaje o incidencia de buenos o malos pagadores para cada caso. A nivel general, se observa que 3,47% del total de casos pertenecen a malos pagadores.

Cuadro 6.
Distribución de variables categóricas según tipo de pagador, enero y diciembre 2023

Covariables	Pagador			% del total	% de buenos	% de malos
	Bueno	Malo	Total			
Pagador	65 457	2 356	67 813	100,00%	96,53%	3,47%
Segmento de Mercado						
Segmento 1 (Pensionados)	41 065	104	41 169	60,71%	99,75%	0,25%
Segmento 2	12 179	876	13 055	19,25%	93,29%	6,71%
Segmento 3	6 292	1 018	7 310	10,78%	86,07%	13,93%
Segmento 4	2 806	140	2 946	4,34%	95,25%	4,75%
Segmento 5	2 330	154	2 484	3,66%	93,80%	6,20%
Segmento 6	599	22	621	0,92%	96,46%	3,54%
Segmento 7	186	42	228	0,34%	81,58%	18,42%
Tipo de Crédito						
Consumo	64 423	2 356	66 779	98,48%	96,47%	3,53%
Vivienda	1 034	0	1 034	1,52%	100,00%	0,00%
Tipo de Garantía						
Hipotecaria	1 179	1	1 180	1,74%	99,92%	0,08%
Fiduciaria	2 695	119	2 814	4,15%	95,77%	4,23%

Covariables	Pagador				% de buenos	% de malos
	Bueno	Malo	Total	% del total		
Sin Garantía	61 583	2 236	63 819	94,11%	96,50%	3,50%
Tenencia de hipotecas						
Sin Hipotecas	63 391	2 343	65 734	96,93%	96,44%	3,56%
Con Hipotecas	2 066	13	2 079	3,07%	99,37%	0,63%
Tenencia de incobrables						
Sin Incobrables	64 422	1 916	66 338	97,82%	97,11%	2,89%
Con Incobrables	1 035	440	1 475	2,18%	70,17%	29,83%
Plan de Inversión						
Personal	23 063	918	23 981	35,36%	96,17%	3,83%
Refundición Externa	21 801	1 026	22 827	33,66%	95,51%	4,49%
Refundición Mixta	9 806	208	10 014	14,77%	97,92%	2,08%
Refundición Interna	9 753	204	9 957	14,68%	97,95%	2,05%
Vivienda	1 034	0	1 034	1,52%	100,00%	0,00%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

La mayoría de las operaciones pertenecen a pagadores pertenecientes al segmento 1 de pensionados, con 60,71% del total de operaciones, pero con una incidencia relativamente baja de malas operaciones de crédito, con solamente 0,34% de operaciones de este segmento.

En cuanto a los restantes segmentos con condición laboral activa, se tiene que el segmento 2, representa el segundo segmento en importancia, pero con 6,71% de malos pagadores para el total del segmento, del mismo modo; el segmento 3 representa el tercer segmento en cantidad de casos, y presentan una incidencia mayor de malos pagadores, con 13,93%. En menor frecuencia se tienen los segmentos 4 y 5, con concentraciones de casos similares de 4,34% y 3,66% respectivamente, mientras que sus niveles de malos pagadores se ubican por encima del total de cartera con 4,75% y 6,20%. Por último, se tienen los segmentos 6 y 7, siendo los segmentos de menor concentración de cartera, el segmento 6 mantiene un nivel de malos pagadores de 3,54%; mientras que el segmento 7 muestra el mayor nivel de incidencia de malos pagadores, con 18,42%.

Sobre las restantes variables, destaca que la gran mayoría de la cartera obedece a créditos de consumo con una concentración de 98,48% del total de operaciones, donde 3,53%

corresponden a operaciones de malos pagadores, mientras que a nivel de créditos de vivienda no se tienen créditos catalogados como malos pagadores. De manera similar, si se observa la distribución por tipo de garantía, la mayor concentración obedece a créditos sin garantía y en menor medida a créditos con garantía fiduciaria e hipotecaria, donde destaca que solamente se tiene un caso con garantía hipotecaria catalogada como mal pagador.

Además, se tiene la variable referente a que, si la operación de crédito pertenece a un deudor que mantenga operaciones de crédito con garantía hipotecaria, y si la operación pertenece a un deudor que ha tenido operaciones pasadas a incobrables en el pasado. En el caso de la primera, se obtiene que la mayoría de las operaciones de crédito pertenecen a deudores que no mantienen hipotecas, los cuales son los que muestran mayor incidencia de malos pagadores con un nivel de 3,56%, mientras que los deudores que si mantienen hipotecas tienen una incidencia de 0,63%. Para el caso de la tenencia de incobrables, se observa que solo 2,18% de las operaciones pertenecen a deudores que tuvieron operaciones incobrables en el pasado, pero tienen una gran incidencia de que esas operaciones lleguen a ser catalogadas como malos pagadores en el futuro, con un resultado de 29,83%.

Por último, se tiene la variable referente al plan de inversión del crédito, referente a si el crédito tiene como objetivo ser de consumo personal, refundir deudas o para solución de vivienda. De esta manera se observa que la mayoría de los créditos otorgados son créditos de refundición en su conjunto, siendo mayor el porcentaje de créditos orientados a refundición externa, le siguen los créditos de consumo personal, y por ultimo los créditos de vivienda. Los créditos orientados a consumo personal y refundición externa muestran los niveles de malos pagadores más elevados 3,83% y 4,49%, y se tiene una incidencia de 0,00% para créditos de vivienda. Por otra parte, el Cuadro 7 muestra la distribución de las variables continuas según el promedio de cada variable para buenos y malos pagadores.

Cuadro 7.
Promedio de variables continuas según tipo de pagador, enero y diciembre 2023

Covariables	Bueno	Malo	Total
Saldo operación	₡8 969 312,77	₡6 125 752,85	₡8 870 520,10
Saldo deudor	₡14 804 119,66	₡9 606 617,02	₡14 623 544,90
Cuota	₡117 335,76	₡81 221,51	₡116 081,05
Cuotas deudor	₡197 223,17	₡128 209,47	₡194 825,45
Diferencia Cuotas	-₡2 951,57	-₡2 398,20	-₡2 932,35
Tasa de interés	12,36%	12,44%	12,36%
Cuotas Pendientes	146,85	129,62	146,25
Amortización	20,40%	21,96%	20,46%
Antigüedad	4,44	4,96	4,46
Días de mora	2,84	128,94	7,22
Mora promedio	2,97	119,89	7,03
Puntaje deudor	1,19	3,36	1,26
Mora máxima	15,15	163,15	20,29
Salario bruto	₡1 281 791,23	₡792 308,25	₡1 264 785,32
Salario neto	₡1 024 952,36	₡645 262,33	₡1 011 760,94
Salario liquido	₡828 112,09	₡520 181,35	₡817 413,78
Capacidad de pago	66,41%	64,51%	66,34%
Cantidad de operaciones	1,00	1,00	1,00

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Sobre las variables relacionadas a saldo de las operaciones se observa una diferencia entre buenos y malos pagadores en cuanto a que los malos pagadores tienden a tener en promedio saldos menores, tanto por operación como por deudor, lo cual se puede explicar debido a que las operaciones de crédito de vivienda, que mantienen saldos superiores a los créditos de consumo, tienden a estar más concentradas en buenos pagadores. Similar al caso anterior, sucede lo mismo al observar el comportamiento promedio de las variables de cuotas por pagar, ya que, tanto a nivel de la variable de cuota del crédito y cuotas totales del deudor, se tiene que los malos pagadores tienden a tener cuotas más bajas.

Por otra parte, la variable de diferencia de cuotas se construye según la diferencia entre la cuota observada al momento de la medición y la cuota de originación del crédito, al obtener un resultado negativo se interpreta que el monto de las cuotas promedio de los créditos han bajado en el tiempo, al dividir entre buenos y malos pagadores se observa que para las operaciones pertenecientes a buenos pagadores la disminución ha sido levemente más

grande. Al mismo tiempo, al observar el comportamiento de la variable de tasa de interés de los créditos no se observan diferencias entre buenos y malos pagadores.

La variable de cuotas pendientes se puede interpretar como un indicador de la maduración del crédito, se observa que los créditos pertenecientes a buenos pagadores tienden a mantener en promedio mayor cantidad de cuotas pendientes, lo que puede estar relacionado a que son créditos más jóvenes.

Similar al caso anterior, la variable de amortización indica el porcentaje del capital que ha sido amortizado al momento de la observación respecto al monto original del crédito, donde se obtiene el resultado de que los malos pagadores han amortizado en promedio un porcentaje levemente mayor, esto por ser créditos más maduros, lo cual se confirma al observar la variable de antigüedad, la cual muestra el tiempo en años que lleva el crédito en cartera desde el momento de otorgamiento hasta la medición, donde se observa que los créditos de malos deudores son en promedio levemente más antiguos.

Posteriormente se tiene las variables orientadas a la medición de morosidad, específicamente días mora de la operación, mora promedio del deudor, puntaje deudor, y mora máxima del deudor. Para todos los casos se cumple con el resultado esperado, en cuanto a que las operaciones pertenecientes a malos pagadores muestran mayores niveles de mora.

Seguidamente se muestran los resultados de las variables referentes al salario del deudor, tanto a nivel bruto, como neto (sin cargas sociales), y a nivel líquido (luego de la aplicación de las cuotas crediticias), además se muestra el indicador de capacidad de pago, el cual refleja el cociente entre salario líquido y salario bruto. Para todos los casos se tiene que los créditos pertenecientes a malos pagadores mantienen menores salarios y liquidez respecto a los créditos pertenecientes a buenos deudores.

Por último, se tiene la variable de cantidad de operaciones promedio que acumula el deudor dueño de la operación, donde se observa que no hay diferencias entre buenos y malos pagadores.

Según los resultados mostrados anteriormente, se observa que las variables de diferencia de cuotas, tasa de interés, amortización, antigüedad y cantidad de operaciones ofrecen poco

poder predictivo, ya que la diferencia de los resultados promedio entre buenos y malos pagadores es relativamente corta. Caso contrario con todas las variables de morosidad, que ofrecen un alto poder predictivo, ya que las diferencias entre los promedios de buenos y malos pagadores son elevadas.

Ahora bien, para las variables restantes, relacionado los resultados cualitativos y cuantitativos anteriores, resulta necesario explorar los resultados de dichas variables según segmento de mercado, ya que, como se mostró anteriormente, el principal segmento de mercado está compuesto por deudores pertenecientes al segmento 1 (ver cuadro 6), los cuales son deudores pensionados, por lo que existe una diferencia sustancial entre los flujos de dinero que perciben respecto a los restantes segmentos de deudores con condición laboral activa, además de que estos últimos tienen un límite a las deducciones que se pueden hacer a su salario por la llamada Ley de Usura (Ley 9859).

De esta manera el Cuadro 8 muestra los promedios de las diferentes variables según la condición de buen pagador, diferenciando a su vez entre deudores pensionados y en condición laboral activa, con el propósito de identificar si las relaciones iniciales observadas en el Cuadro 7 se mantienen entre estas dos poblaciones.

Cuadro 8.
Promedio de variables continuas moderadamente relevantes por condición laboral del deudor según tipo de pagador, enero y diciembre 2023

Covariables	Bueno	Malo	Total
Saldo operación	₪8 969 312,77	₪6 125 752,85	₪8 870 520,10
Pensionados	₪8 594 642,58	₪9 836 861,57	₪8 597 780,64
Activos	₪9 600 086,44	₪5 954 369,50	₪9 291 943,72
Saldo deudor	₪14 804 119,66	₪9 606 617,02	₪14 623 544,90
Pensionados	₪15 137 923,04	₪16 337 281,42	₪15 140 952,83
Activos	₪14 242 147,05	₪9 295 787,04	₪13 824 071,58
Cuota	₪117 335,76	₪81 221,51	₪116 081,05
Pensionados	₪114 609,11	₪137 384,98	₪114 666,65
Activos	₪121 926,18	₪78 627,81	₪118 266,52
Cuotas deudor	₪197 223,17	₪128 209,47	₪194 825,45
Pensionados	₪205 779,08	₪224 915,02	₪205 827,42
Activos	₪182 818,91	₪123 743,49	₪177 825,75
Cuotas pendientes	146,85	129,62	146,25
Pensionados	141,13	145,96	141,15
Activos	156,48	128,86	154,15

Covariables	Bueno	Malo	Total
Salario Bruto	¢1 281 791,23	¢792 308,25	¢1 264 785,32
Pensionados	¢1 280 832,00	¢1 175 941,53	¢1 280 567,03
Activos	¢1 283 406,14	¢774 591,61	¢1 240 400,20
Salario Neto	¢1 024 952,36	¢645 262,33	¢1 011 760,94
Pensionados	¢1 018 906,45	¢920 924,39	¢1 018 658,93
Activos	¢1 035 130,90	¢632 531,94	¢1 001 102,49
Salario Liquido	¢828 112,09	¢520 181,35	¢817 413,78
Pensionados	¢813 174,99	¢702 142,32	¢812 894,50
Activos	¢853 259,36	¢511 778,18	¢824 396,74
Capacidad de Pago	66,41%	64,51%	66,34%
Pensionados	66,85%	60,04%	66,83%
Activos	65,67%	64,72%	65,59%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Iniciando con la variable de saldo de operación, se tiene que a pesar de que inicialmente se encontraba que las operaciones de malos pagadores tenían saldos más bajos que los buenos pagadores, al diferenciar por segmentos, se observa que la relación se mantiene para deudores activos, pero se invierte para deudores pensionados, donde a su vez la diferencia entre buenos y malos pagadores se reduce, lo anterior se repite para las variables de saldo deudor, cuota, cuotas deudor, y cuotas pendientes.

Para el caso de las variables salariales se observa que la relación inicial de una relación de mayor salario y liquidez con el grupo de buenos pagadores se mantiene al separar entre pagadores activos y pensionados, sin embargo, la diferencia dentro del grupo de pagadores pensionados se reduce sustancialmente, lo cual indica un poder predictivo mucho menor para este grupo respecto al poder predictivo del grupo de pagadores activos.

4.2. MODELOS DE APRENDIZAJE NO SUPERVISADO

Por otro lado, en esta sección se muestra el análisis realizado según los modelos de aprendizaje no supervisado como lo son los modelos de análisis descriptivo es el análisis de componentes principales y el análisis de “K” medias.

4.2.1. Análisis de Componentes Principales

Para este caso, dado que las metodologías señaladas se basan en variables cuantitativas, y las variables que aportan mayor poder predictivo de este tipo son las variables de morosidad

señaladas anteriormente, se utilizan estas variables con el propósito de identificar si es posible crear una variable que resuma los indicadores de morosidad en una variable alterna, que pueda ser implementada en el modelo de clasificación final.

De esta manera, el Cuadro 9 muestra la matriz de correlaciones de las variables de morosidad, donde se tiene que todas las variables presentan altas correlaciones entre sí, el cual es un resultado esperado.

Cuadro 9.
Matriz de correlaciones variables de morosidad, enero y diciembre 2023

Covariables	Días mora	Mora Promedio	Puntaje deudor	Mora máxima
Días mora	1,00	0,96	0,65	0,65
Mora Promedio	0,96	1,00	0,67	0,68
Puntaje deudor	0,65	0,67	1,00	0,76
Mora máxima	0,65	0,68	0,76	1,00

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Al realizar el análisis de componentes principales, se determina por medio de Cuadro 10, que se puede explicar hasta 80% de la variancia solamente con 1 componente, y 93% al utilizar dos componentes.

Cuadro 10.
Análisis ACP variables de morosidad, enero y diciembre 2023

Variables	PC1	PC2	PC3	PC4
Desviación estándar	1,79	0,73	0,49	0,21
Proporción de variancia	0,80	0,13	0,06	0,01
Proporción acumulada	0,80	0,93	0,99	1,00

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Al examinar las cargas factoriales sobre los componentes es fácil identificar que todos muestran cargas factoriales sobre la primera dimensión.

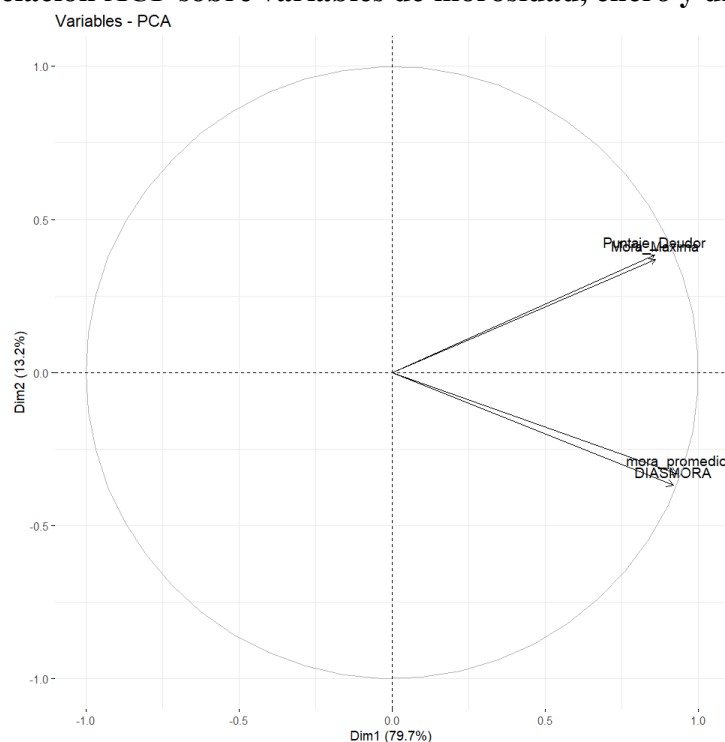
Cuadro 11.
Cargas factoriales ACP variables de morosidad, enero y diciembre 2023

Covariables	PC1	PC2	PC3	PC4
Días mora	-0,51	-0,51	-0,01	0,69
Mora Promedio	-0,52	-0,45	-0,01	-0,72
Puntaje deudor	-0,48	0,53	-0,70	0,02
Mora máxima	-0,48	0,51	0,71	0,02

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Por su parte, el Grafico 5 muestra el resultado del círculo de correlaciones según los resultados del ACP.

Gráfico 5.
Círculo de correlación ACP sobre variables de morosidad, enero y diciembre de 2023



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Es claro observar que todas las variables mantienen una relación alta con la dimensión 1, y que las variables que resumen morosidad histórica, como lo son las variables de puntaje deudor y mora máxima están muy correlacionadas, lo mismo sucede con las variables de días de mora y morosidad promedio, las cuales se miden al cierre de cada mes.

4.2.2. Análisis de “K” medias

Seguidamente, se desarrolla un análisis de agrupamiento de las observaciones por medio del análisis “K” medias, donde se define la construcción de 3 grupos que permitan identificar diferentes niveles de morosidad.

El Cuadro 12 muestra los centros de los grupos construidos, donde es fácil observar que las tres agrupaciones se pueden diferenciar claramente según nivel de morosidad, por lo que se puede definir el clúster 1 como un clúster donde se ubican los casos con baja mora actual e histórica, el clúster 2 agrupa los casos con un nivel medio, y el clúster 3 agrupa los casos con los niveles más altos de morosidad.

Cuadro 12.
Centros de clúster K medias para variables de morosidad, enero y diciembre de 2023

Grupos	Días mora	Mora Promedio	Puntaje deudor	Mora máxima
1	0,50	0,57	2,28	1,79
2	-0,19	-0,19	-0,29	-0,24
3	4,86	4,82	3,66	3,43

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Al comparar los grupos generados con la variable explicativa, se obtienen los resultados contenidos en el Cuadro 13.

Cuadro 13.
Clúster por K medias para variables de morosidad según tipo de pagador, enero y diciembre de 2023

Clúster	Buen Pagador	Mal Pagador	Total	% del total	% malos
Bajo	60 867	326	61 193	90,24%	0,53%
Medio	4 166	556	4 722	6,96%	11,77%
Alto	424	1 474	1 898	2,80%	77,66%
Total	65 457	2 356	67 813	100,00%	3,47%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se observa que los grupos ofrecen una diferenciación clara de los casos según buenos y malos pagadores, ya que el clúster bajo, muestra una incidencia de mora de 0,5%, con una concentración de 90,24% del total de casos, el clúster medio presenta una incidencia de

11,77% con una concentración de 6,96%, mientras que el clúster de mora alta tiene la mayor incidencia de mora, con 77,66% y una concentración de 2,80% del total de casos.

Según lo anterior, se decide considerar los grupos construidos como una nueva posible variable explicativa a la que se llamará “Grupo de mora” y será evaluada en la construcción de los modelos de aprendizaje supervisado que se mostraran en la siguiente sección.

4.3. MODELOS DE APRENDIZAJE SUPERVISADO

Según los resultados anteriores, se realizó la selección de variables de análisis para la fase de modelado, donde se consideró tanto los resultados del análisis descriptivo para evaluar el poder predictivo de las variables, como el análisis de los modelos de aprendizaje no supervisado, las variables seleccionadas y los criterios que apoyan su selección se muestran en el Cuadro 14.

Cuadro 14.
Selección de variables para fase de modelado

Covariables	Selección	Criterio
Cualitativas		
Segmento de Mercado	Si	Diferencias altas entre categorías
Tipo de Crédito	No	Pocas observaciones en categoría vivienda, y efectos mitigadores en el proceso de pérdida esperada para estos casos
Tipo de Garantía	No	Pocas observaciones en categoría vivienda, y efectos mitigadores en el proceso de pérdida esperada para estos casos
Tenencia de hipotecas	No	Moderado poder predictivo, concentrado solo en una categoría
Tenencia de incobrables	Si	Diferencias altas entre categorías
Plan de Inversión	Si	Diferencias bajas entre categorías
Grupo de Mora	No	La inclusión de esta variable no aporta en mejorar los resultados
Cuantitativas		
Saldo operación	No	Poder predictivo no consistente entre pagadores activos y pensionados
Saldo deudor	No	Poder predictivo no consistente entre pagadores activos y pensionados

Covariables	Selección	Criterio
Cuota	No	Poder predictivo no consistente entre pagadores activos y pensionados
Cuotas deudor	No	Poder predictivo no consistente entre pagadores activos y pensionados
Diferencia Cuotas	No	Diferencias bajas entre buenos y malos pagadores
Tasa de interés	No	Sin poder predictivo
Cuotas Pendientes	No	Poder predictivo no consistente entre pagadores activos y pensionados
Amortización	No	Diferencias bajas entre buenos y malos pagadores
Antigüedad	No	Diferencias bajas entre buenos y malos pagadores
Días de mora	Si	Diferencias altas y consistentes entre buenos y malos pagadores
Mora promedio	Si	Diferencias altas y consistentes entre buenos y malos pagadores
Puntaje deudor	Si	Diferencias altas y consistentes entre buenos y malos pagadores
Mora máxima	Si	La inclusión de esta variable no aporta en mejorar los resultados
Salario bruto	No	Poder predictivo no consistente entre pagadores activos y pensionados
Salario neto	No	Poder predictivo no consistente entre pagadores activos y pensionados
Salario liquido	No	Poder predictivo no consistente entre pagadores activos y pensionados
Capacidad de pago	No	Poder predictivo no consistente entre pagadores activos y pensionados
Cantidad de operaciones	No	Diferencias bajas entre buenos y malos pagadores

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

De esta manera, se obtiene llega a una base de datos de 67 813 observaciones, 5 variables explicativas y la variable respuesta definida como “Pagador”, las variables seleccionadas son las siguientes:

- Segmento de mercado

- Tenencia de incobrables
- Días de mora
- Mora promedio
- Puntaje deudor

Cabe resaltar que, pese a que en el análisis descriptivo se observa una diferencia entre buenos y malos deudores para la variable mora máxima, se decidió excluirla debido a que en modelos preliminares se determinó que la presencia de dicha variable no aportaba en mejorar los resultados de eficiencia de los modelos, lo cual también sucede con la variable de grupo de mora, creada bajo la metodología de análisis de clúster k medias⁷. A su vez, pese a la alta correlación que presentan las variables de días mora con mora promedio, se decide mantenerla dentro de las variables explicativas ya que dicha variable aporta a diferenciar entre deudores que mantienen más de una operación de crédito, pero con diferentes niveles de mora entre ellas.

Por otra parte, a pesar de que se tiene una incidencia de mora muy baja en créditos de vivienda, que por defecto tienen garantía hipotecaria, y créditos de consumo con la misma garantía, se decidió no considerar dichas variables debido a la poca cantidad de observaciones en la base de datos total, además de que dichos casos ya cuentan con efectos mitigadores para el cálculo de riesgo de crédito en el proceso de pérdida esperada.

A partir de lo anterior, según lo mencionado en el capítulo metodológico, se procedió a realizar una fase previa de calibración con el propósito de seleccionar, los mejores parámetros de ciertos modelos de interés, con el propósito de maximizar la eficiencia de los modelos a probar (ver anexo 2).

El Cuadro 15 muestra los resultados de los indicadores de clasificación de los modelos generados con los mejores parámetros, dichos indicadores corresponden a los promedios de las iteraciones llevadas a cabo bajo el enfoque de validación cruzada.

⁷ En la sección de anexos se muestran los resultados de las pruebas realizadas en el Anexo 1.

Cuadro 15.
Resultados de evaluación de modelos calibrados, enero y diciembre de 2023

Modelos	Parámetros	Precisión global	Precisión buenos pagadores	Precisión malos pagadores
Regresión logística	Logit	98,05%	99,46%	58,93%
Bosques aleatorios	Cantidad de árboles = 1000 (variables por interacción = 13)	98,06%	99,24%	65,31%
Arboles de decisión	Mínimo número individuos en cada nodo = 500	98,18%	99,35%	65,59%
Máquinas de soporte vectorial (SVM)	Kernel: “linear”	98,19%	99,35%	65,85%
XG boost	Booster: “gbtree” (Cantidad de árboles = 1000)	98,13%	99,26%	66,75%
Ada boost	Booster: “gentle” (Cantidad de árboles = 300)	98,15%	99,27%	67,12%
K vecinos	Kernel: “inv”	98,17%	99,25%	68,12%
Análisis discriminante	NA	97,95%	98,88%	72,10%
Bayes	NA	95,77%	96,27%	81,95%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se obtiene una diferencia mínima en los indicadores de precisión global, donde los mejores modelos son k vecinos, máquinas de soporte vectorial, y arboles de decisión, con niveles entre 98,17% y 98,18%. De manera similar, se obtiene diferencias mínimas en cuanto a la clasificación de operaciones de crédito de buenos pagadores, donde los resultados arrojan que el modelo de regresión logística obtiene el mejor resultado, con una clasificación de 99,46%. Por último, la clasificación de operaciones de crédito de malos pagadores muestra

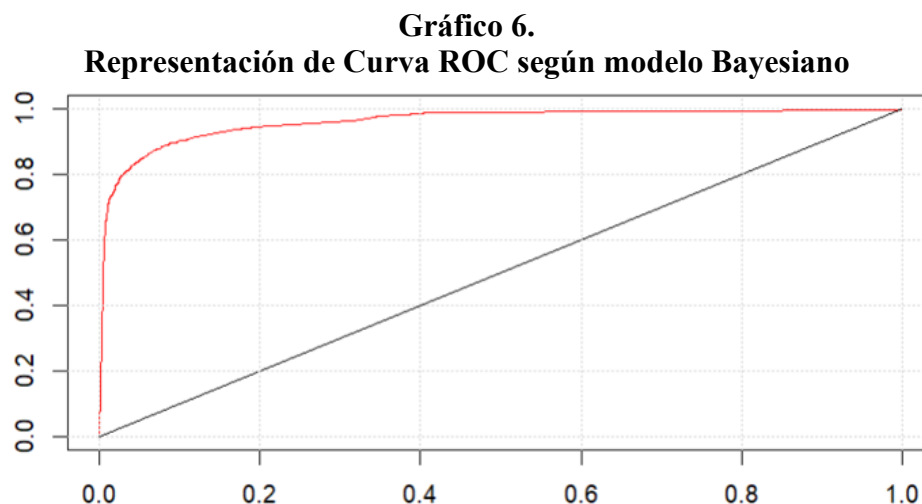
diferencias más marcadas, donde el mejor resultado se obtiene en el modelo de Bayes, con una clasificación de 81,95%.

Es importante mencionar que, salvo el caso del modelo de regresión logística, todos los demás cumplen con los criterios mínimos de aceptación para considerar que los modelos son adecuados⁸, ya que obtiene un nivel de clasificación de operaciones de malos pagadores superior a 60,00%, y de clasificación global y de buenas operaciones superior a 85,00%.

De este modo, desde un enfoque de riesgo de crédito y dada la necesidad principal de identificar a los malos pagadores por encima de los buenos pagadores, se determina que el modelo que ofrece los mejores resultados es el modelo de Bayes.

Según el resultado del modelo Bayesiano, se calcula el indicador F1 Score sobre la totalidad de la base de datos, lo cual arroja un resultado de 0,58, lo que indica una capacidad medianamente aceptable de clasificación del modelo.

Adicionalmente, el Grafico 6 muestra la Curva ROC correspondiente, de donde se puede calcular el indicador área bajo la curva (AUC por sus siglas en ingles) de 0,96, lo que indica una alta capacidad del modelo para distinguir entre las dos clases.

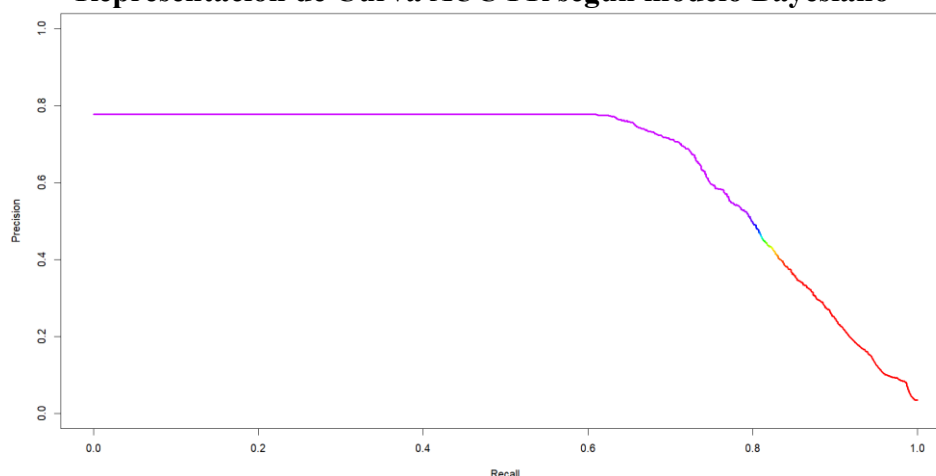


Fuente: elaboración propia a partir de datos de crédito de JUPEMA

⁸ Ver Capítulo III, sección de plan de pruebas.

En complemento a los anterior, para casos desbalanceados un indicador alternativo al AUC es el AUC PR, el cual se basa en las medidas de precisión-recuperación de la matriz de confusión para todas las diferentes probabilidades de corte. De esta manera, el Grafico 7 muestra la curva AUC PR, la cual arroja un indicador de área bajo la curva de 0,65; lo cual puede interpretarse como un resultado de clasificación bueno para el modelo elegido.

Gráfico 7.
Representación de Curva AUC PR según modelo Bayesiano



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Además, se realizó el análisis de probabilidad de corte, donde se determinó que al utilizar una probabilidad diferente a 0,50 no mejora sustancialmente los resultados (ver anexo 3).

4.4. VALIDACIÓN DE RESULTADOS Y CONSTRUCCIÓN DE CATEGORÍAS DE RIESGO

Según lo mencionado en el Capítulo III sobre la metodología a seguir, una vez elegido el mejor modelo de clasificación, es necesario validar sus resultados con datos nuevos. Para este paso se utilizan todos los datos fuera de muestra disponibles para el año 2023 en el periodo de febrero a noviembre, siendo un total de 338.380 observaciones, y se analiza la distribución de las variables explicativas según la probabilidad de incumplimiento predicha.

Por otro lado, un factor para considerar en esta fase es que uno de los requerimientos dentro del proceso de riesgo de crédito es poder generar categorías de riesgo, en este caso por operación de crédito a partir de las probabilidades de incumplimiento predichas. Para lo

anterior, es necesario generar cortes en la distribución de probabilidades predichas que permitan categorizar según nivel de riesgo a las operaciones.

El Cuadro 16 muestra la distribución del resultado de las probabilidades predichas sobre la base de datos nuevos según deciles, se observa una alta concentración de valores con probabilidades menores a 0,10, con alrededor de 94,31% de los casos, seguida de una concentración de valores con probabilidades mayores a 0,90, con 4,21% de los casos, y concentraciones menores a 1,00% entre esas probabilidades. Este resultado es consistente con el resultado esperado, ya que la cartera de crédito se mantuvo con un bajo nivel de deterioro, donde en promedio 94,11% del total de cartera se mantuvo al día con sus obligaciones de crédito durante el año 2023.

Cuadro 16.
Distribución de probabilidades predichas sobre base de datos nuevos, febrero a noviembre de 2023

Probabilidades	Frecuencia	Porcentaje
Menor a 0,10	319 125	94,31%
Entre 0,10 y 0,20	1 313	0,39%
Entre 0,20 y 0,30	913	0,27%
Entre 0,30 y 0,40	360	0,11%
Entre 0,40 y 0,50	397	0,12%
Entre 0,50 y 0,60	197	0,06%
Entre 0,60 y 0,70	310	0,09%
Entre 0,70 y 0,80	489	0,14%
Entre 0,80 y 0,90	1 015	0,30%
Mayor a 0,90	14 261	4,21%
Total	338 380	100,00%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Es por la acumulación anterior, que se procede a generar cortes explorando la distribución por cuartiles, el Cuadro 17 muestra el resultado de los cuartiles por encima del cuartil 90.

Cuadro 17.
Distribución probabilidades predichas según cuartiles, febrero a noviembre de 2023

Cuartiles	Probabilidades
Cuartil 90	0,05%
Cuartil 92,5	0,32%
Cuartil 95	33,13%
Cuartil 97,5	99,99%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se obtiene que el 90,00% de las observaciones tienen una probabilidad de incumplimiento de 0,05%, mientras que 92,50% se ubican con una probabilidad levemente superior, de 0,32%. Al aumentar al cuartil 95, se obtiene que el 95,00% de las observaciones se ubican con una probabilidad por debajo de 33,13%, por último, el percentil 97,50 arroja que el 97,5% se ubican con una probabilidad de 99,99%, dejando a 2,50% de observaciones por encima de ese nivel.

Según lo anterior, y considerando que, bajo el esquema de la metodología de calificación de riesgo actual, la concentración de cartera de crédito en la categoría de menor riesgo se ubica cerca de 90,00%, y la concentración de cartera en la categoría de mayor riesgo se ubica cerca de 5,00%. Además, tomando en cuenta que, según el modelo de clasificación, a partir una probabilidad de 50,00% es posible clasificar a una operación de crédito como perteneciente a un mal pagador y permite obtener una clasificación adecuada de buenos y malos pagadores, se define utilizar este corte por ser razonable con el apetito al riesgo de la institución, en el sentido de que deudores con una probabilidad de incumplimiento mayor a 50,00%, no serían sujetos a nuevos créditos.

De esta manera, se define utilizar los cortes mostrados en el Cuadro 18, donde se muestran 4 cortes que permitan construir hasta cinco categorías de riesgo, las cuales para efectos prácticos se nombrarán de manera alfabética de menor a mayor nivel de riesgo.

Cuadro 18.

Definición de categorías de riesgo según probabilidades de incumplimiento, febrero a noviembre de 2023

Categorías de Riesgo	Probabilidades
Categoría A	Menor o igual 0,05%
Categoría B	Entre 0,05% y 33,13%
Categoría C	Entre 33,13% y 50,00%
Categoría D	Entre 50,00% y 99,99%
Categoría E	Mayor a 99,99%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Según los cortes anteriores se generan las respectivas categorías de riesgo. El Cuadro 19 muestra la distribución de las operaciones de la base de datos de validación respecto a las

categorías de riesgo, además se observa el porcentaje de casos que efectivamente fueron pasados a incobrables un año posterior al momento de observación.

Cuadro 19.
Distribución de operaciones en datos de validación según categorías de riesgo construidas, febrero a noviembre de 2023

Categoría	Operaciones	Porcentaje	% Casos Incobrables
A	304 807	90,08%	0,29%
B	16 654	4,92%	4,22%
C	674	0,19%	5,56%
D	7 792	2,30%	11,36%
E	8 480	2,51%	41,65%
Total	338 380	100,00%	1,79%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

En primera instancia, se obtiene que 90,08% de las operaciones se consideran como buenas operaciones, asignándolas en la categoría A, la categoría de menor riesgo. Luego 4,92% se asignan a la categoría B, en categoría C se acumula 0,19%, mientras que, en las categorías de mayor riesgo, 2,30% se ubican en categoría D y 2,51% en la categoría E.

Según las categorías de riesgo construidas, el Cuadro 20 muestra la distribución de las variables cualitativas según las diferentes categorías. Por otro lado, al observar la proporción de casos que efectivamente fueron pasados a incobrables un año después por categoría de riesgo, se observa que efectivamente se tiene que entre mayor es la categoría de riesgo se aumenta la incidencia de casos a incobrables, cumpliendo con el resultado esperado.

Cuadro 20.
Distribución de variables cualitativas según categorías de riesgo, febrero a noviembre de 2023

Covariables	Categorías de Riesgo					Total
	A	B	C	D	E	
Segmento de Mercado						
Segmento 1 (Pensionados)	98,93%	0,50%	0,03%	0,34%	0,21%	100,00%
Segmento 2	75,13%	12,02%	0,43%	5,97%	6,46%	100,00%
Segmento 3	66,05%	16,94%	0,66%	7,03%	9,32%	100,00%
Segmento 4	89,00%	6,36%	0,12%	2,33%	2,20%	100,00%
Segmento 5	89,13%	6,63%	0,28%	2,32%	1,65%	100,00%
Segmento 6	85,25%	6,03%	0,00%	4,97%	3,75%	100,00%
Segmento 7	47,33%	26,04%	3,76%	11,88%	10,99%	100,00%
Tenencia de Incobrables						
Con Incobrables	16,16%	29,28%	2,16%	22,84%	29,57%	100,00%
Sin Incobrables	91,58%	4,43%	0,15%	1,89%	1,96%	100,00%
Total	90,08%	4,92%	0,19%	2,30%	2,51%	100,00%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se observa que, para el caso de segmento de mercado, el segmento perteneciente a deudores pensionados concentra el mayor porcentaje de deudores en categoría A, siendo esta la categoría de menor riesgo, y a su vez, concentra el menor porcentaje de deudores en categoría E, la cual es la categoría de mayor riesgo, es necesario recordar que el segmento de pensionados presenta la menor incidencia de operaciones pertenecientes a malos pagadores, por lo que el resultado es acorde a lo esperado.

De manera análoga, los segmentos 4, 5, y 6, que son los segmentos con los niveles de deterioro más bajos después del segmento de pensionados, muestran un nivel menor de concentración en categoría A, pero mayor en categoría E y las categorías intermedias, mientras que los segmentos 2, 3, y 7, muestran los mayores niveles de concentración en categorías de mayor riesgo, lo cual es acorde por ser los segmentos con mayor deterioro.

Del mismo modo, la variable de tenencia de operaciones incobrables cumple con el resultado esperado, ya que la categoría de tenencia de operaciones incobrables muestra una concentración marcadamente mayor en las categorías de mayor riesgo respecto a la categoría contraria.

En cuanto a las variables cuantitativas, el Cuadro 21 muestra el promedio de cada variable según las diferentes categorías de riesgo.

Cuadro 21.
Resultados promedio de variables cuantitativas según categorías de riesgo, febrero a noviembre de 2023

Categorías de Riesgo	Días de mora	Mora promedio deudor	Puntaje Deudor
Categoría A	0,36	0,44	1,04
Categoría B	14,09	15,16	2,17
Categoría C	19,67	20,59	2,95
Categoría D	28,99	31,18	3,34
Categoría E	122,62	115,47	3,85
Total	4,80	4,79	1,22

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se cumple que, a mayor nivel de categoría de riesgo, se tiene un promedio más elevado, específicamente se tiene un promedio de 0,36 días de mora para los casos en categoría A, contra 122,62 días de mora para casos en categoría E. Para el caso de mora promedio del deudor se pasa de un nivel de 0,44 días, en categoría A, hasta 115,47 en categoría E, y para el caso de la variable de puntaje deudor, se pasa de un nivel promedio de 1,04 en categoría A, hasta 3,85 en categoría E⁹.

Según los resultados anteriores, se considera que tanto el modelo de clasificación utilizado para generar las probabilidades de incumplimiento, así como las categorías de riesgo generadas a partir de dicho resultado son razonables según el nivel de deterioro mostrado por la cartera y las variables explicativas.

Además, resulta importante mencionar que en lo que respecta a la construcción de las categorías de riesgo a partir de las probabilidades de incumplimiento, a pesar de que los cortes utilizados podrían parecer extremos, esto se debe a la polarización de las probabilidades, lo que a su vez se explica por la polarización que se genera al utilizar las variables relacionadas con días de atraso, ya que en promedio más del 95,00% de las operaciones se mantienen al día de casos al día. Dicho fenómeno podría generar inestabilidad en el comportamiento de las categorías, sin embargo, el uso de variables categóricas no

⁹ En la sección de anexos se muestra el Anexo 4 con los análisis gráficos de las distribuciones.

ligadas directamente con los días de atraso de las operaciones permite mitigar este posible fenómeno, lo anterior se puede corroborar al observar el comportamiento de las categorías de riesgo según las variables explicativas para el año 2024, donde se obtiene un comportamiento similar al año 2023 (ver anexo 5).

4.5. RESULTADOS DE CÁLCULO DE PERDIDA ESPERADA

En esta sección se presentan los resultados de la aplicación del mejor modelo obtenido y las categorías de riesgo, dentro del proceso de cálculo de pérdida esperada de la cartera de crédito, con el propósito de identificar si el esquema de asignación de probabilidades de incumplimiento según un modelo multivariado es viable y aporta un mejor resultado en función del deterioro real de la cartera.

Según el esquema mostrado en el capítulo metodológico, para lograr lo anterior, se aplican tanto el mejor modelo obtenido y los cortes de probabilidades de incumplimiento para las categorías de riesgo, en una base de datos completamente nueva, esto con el objetivo de simular lo más cercano posible la situación real, ya que normalmente los resultados de un año calendario definen los parámetros a utilizar en el siguiente año, de esta manera se toman las observaciones correspondientes al año 2024, donde se tiene una base de datos de 422.728 registros.

Para este punto, es necesario recordar el criterio metodológico que sigue el cálculo de la Perdida Esperada de la cartera, el cual se mostró en la sección 2.4, por medio de la ecuación 1, donde el resultado de la Perdida Esperada de la cartera depende de la relación de la probabilidad de incumplimiento, la exposición o pérdida potencial de la operación, y la pérdida dado el incumplimiento.

De este modo, para realizar el recalcu del monto de Perdida Esperada, se utilizan las nuevas probabilidades de incumplimiento por operación de crédito y se mantienen constantes los valores de exposición y pérdida por incumplimiento, y finalmente se suma el resultado obtenido por mes.

El Cuadro 22 muestra la comparación del resultado obtenido respecto al monto de pérdida esperada registrada para el año 2024.

Cuadro 22.
Resultados y comparación de pérdida esperada de la cartera por mes, enero a diciembre de 2024 (montos en colones)

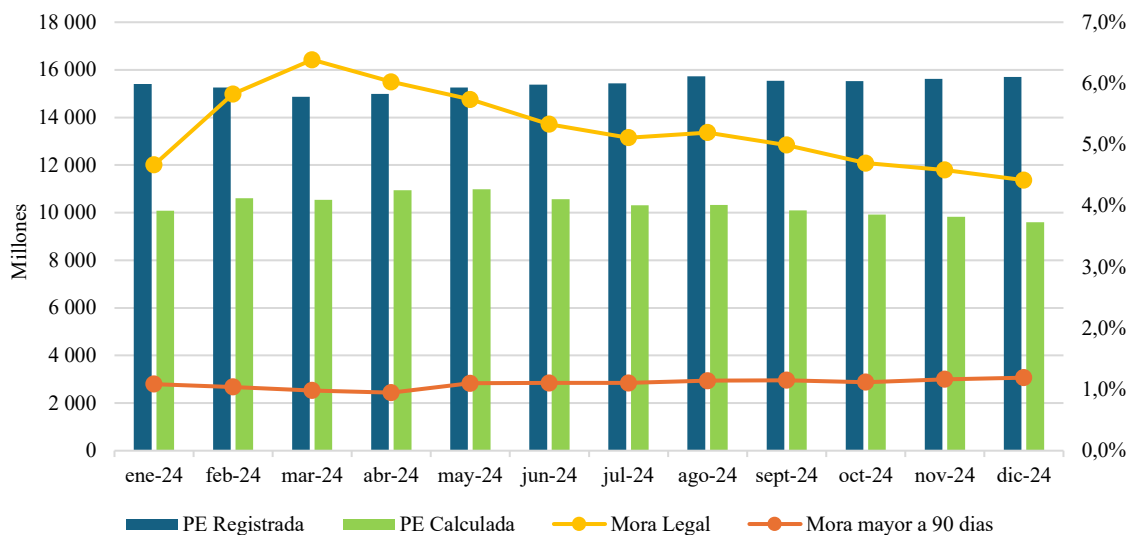
Periodo	Pérdida Esperada		Cambio mensual		Diferencia
	Registrado	Calculado	Registrado	Calculado	
ene-24	15 406 600 410	10 080 761 543			5 325 838 867
feb-24	15 262 392 456	10 606 970 467	-144 207 954	526 208 923	4 655 421 989
mar-24	14 863 698 059	10 541 853 030	-398 694 397	-65 117 437	4 321 845 029
abr-24	14 992 311 526	10 941 577 954	128 613 467	399 724 924	4 050 733 573
may-24	15 258 929 812	10 985 608 701	266 618 285	44 030 747	4 273 321 111
jun-24	15 374 927 439	10 570 836 228	115 997 627	-414 772 473	4 804 091 210
jul-24	15 432 817 708	10 311 310 720	57 890 269	-259 525 508	5 121 506 988
ago-24	15 729 467 057	10 321 736 321	296 649 349	10 425 601	5 407 730 736
sept-24	15 542 856 272	10 097 276 058	-186 610 785	-224 460 263	5 445 580 214
oct-24	15 530 475 220	9 923 805 056	-12 381 052	-173 471 003	5 606 670 165
nov-24	15 621 450 065	9 825 865 051	90 974 845	-97 940 004	5 795 585 014
dic-24	15 707 722 140	9 602 691 622	86 272 075	-223 173 429	6 105 030 518
Promedio	15 392 458 887	10 339 048 292	27 374 703	-43 460 902	5 076 112 951

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se obtiene que el esquema de definición según el modelo multivariado arroja un resultado de disminución en el monto de pérdida esperada, que en promedio se ubica en ₡5.076 millones, a su vez se obtiene que, a diferencia del monto registrado, donde se tiene una tendencia de aumento en la pérdida esperada de la cartera, el monto calculado muestra una tendencia de disminución, junto con una mayor variabilidad mensual con cambios mensuales que en promedio son mayores respecto al dato registrado.

Para tratar de explicar el comportamiento del monto de pérdida esperada bajo ambos esquemas, el Grafico 7 muestra la comparación con los indicadores de mora de cartera, tanto a nivel de mora legal o a partir de un día de atraso, como mayor a 90 días.

Gráfico 8.
Comparación monto de pérdida esperada registrada y calculada con indicadores de mora de la cartera, enero - diciembre 2024 (montos en millones de colones)



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Se observa que la metodología calculada tiene una mayor relación con el indicador de mora legal, con una correlación de 0,87, siendo una correlación alta y positiva, lo que indica que tienen una relación directa, por lo que en la medida de que aumenta la mora general de la cartera, aumenta el monto de pérdida esperada. Al contrario, la relación con el dato de pérdida esperada registrado arroja una correlación de -0,88 con la mora legal, lo que indica que entre menor sea el nivel de mora de la cartera, mayor es el nivel de pérdida esperada.

Al realizar el mismo análisis con el indicador de mora mayor a 90 días, se obtiene un resultado inverso, ya que la correlación del método calculado con el indicador de mora legal es de -0,72, mientras que la correlación con el método registrado es de 0,93, lo que indica que el monto de pérdida esperada registrado aumenta en la medida de que aumenta la mora mayor a 90 días, mientras que el monto de pérdida esperada calculada disminuye.

Lo anterior se explica debido a que la metodología con las probabilidades generadas con el modelo desarrollado, dependen en gran medida de los días de atraso, aumentando más rápidamente la probabilidad de incumplimiento en la medida de que aumenta el nivel de mora, y viceversa, reaccionando más a corto plazo, mientras que en la metodología actual, el deterioro se refleja más lentamente y requiere alcanzar una cantidad elevada de días de atraso

y mantenerse en ese nivel más tiempo para reflejar probabilidades de incumplimiento altas, reaccionando más a mediano o largo plazo.

Por último, es necesario comparar el nivel de cobertura que arroja el resultado calculado, ya que a nivel institucional se tiene definido según apetito de riesgo, que el monto de pérdida esperada debe cubrir un mínimo de 2 veces el saldo de la cartera con mora mayor a 90 días, lo anterior permite establecer si la definición de la probabilidad de incumplimiento según métodos estadísticos multivariados resulta viable a nivel institucional para poder asumir un mayor nivel de riesgo, pero sin superar el límite establecido o los mínimos tolerables.

El Cuadro 23 muestra la comparación de los indicadores de cobertura, donde se observa como la cobertura que arroja el monto de pérdida esperada calculado es menor que el registrado, pero se ubica por encima del objetivo de cobertura establecido de 2 veces, siendo en promedio 2,90 veces contra 4,31 veces del monto registrado.

Cuadro 23.
Comparación indicadores de cobertura registrada y calculada respecto a saldo de cartera con más de 90 días de atraso, enero - diciembre 2024 (montos en colones)

Periodo	Pérdida Esperada		Saldo con mora mayor a 90 días	Cobertura	
	Registrada	Calculada		Registrada	Calculada
ene-24	15 406 600 410	10 080 761 543	3 561 830 510	4,33	2,83
feb-24	15 262 392 456	10 606 970 467	3 404 955 128	4,48	3,12
mar-24	14 863 698 059	10 541 853 030	3 206 284 707	4,64	3,29
abr-24	14 992 311 526	10 941 577 954	3 094 684 144	4,84	3,54
may-24	15 258 929 812	10 985 608 701	3 588 847 442	4,25	3,06
jun-24	15 374 927 439	10 570 836 228	3 604 944 123	4,26	2,93
jul-24	15 432 817 708	10 311 310 720	3 605 664 278	4,28	2,86
ago-24	15 729 467 057	10 321 736 321	3 723 137 394	4,22	2,77
sept-24	15 542 856 272	10 097 276 058	3 759 787 410	4,13	2,69
oct-24	15 530 475 220	9 923 805 056	3 671 315 991	4,23	2,70
nov-24	15 621 450 065	9 825 865 051	3 825 338 177	4,08	2,57
dic-24	15 707 722 140	9 602 691 622	3 908 067 752	4,02	2,46
Promedio	15 392 458 887	10 339 048 292	3 579 571 421	4,31	2,90

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

CAPÍTULO V. CONCLUSIONES

Con la información existente proporcionada por JUPEMA sobre datos históricos de la cartera de crédito del RCC para los años 2023 y 2024, correspondiente a 826 565 observaciones y 29 variables de estudio, y mediante la aplicación de técnicas estadísticas de análisis de aprendizaje supervisado y no supervisado, se llega a las siguientes conclusiones.

Según la distribución de las observaciones considerando la variable respuesta, se obtiene que 3,47% de las observaciones corresponden a operaciones de malos pagadores, lo que establece un ejercicio de construcción de un *score* de crédito con un desbalance en la variable lo que es común en un contexto financiero.

De acuerdo con las variables explicativas, se encuentra que el segmento de mercado ofrece un importante poder predictivo, lo cual se explica debido a que la dinámica laboral a nivel de nombramientos, pago de recargos y nivel salarial difiere mucho entre los diferentes tipos de instituciones educativas, y difieren en gran medida entre población laboralmente activa y pensionados, para el segmento de pensionados la incidencia de malos pagadores es de 0,25%, mientras que para los segmentos con condición laboral activa ronda entre 3,54% y 18,42%.

Otra de las variables a resaltar es la tenencia de operaciones incobrables. Si al momento de la medición el deudor ha tenido operaciones incobrables en el pasado, se aumenta en gran medida la posibilidad de que las restantes operaciones que aún se mantienen activas, también lleguen a ser pasadas como incobrables. Se observa que las operaciones cuyos dueños han sufrido pases a incobrables tienen una incidencia de malos pagadores de 29,83%, siendo muy superior al caso contrario donde no se han dado pases a incobrables, con 2,89%.

Sobre las variables cuantitativas (ver Cuadro 7), existen diferencias en cuanto a que los malos pagadores muestran una tendencia a tener saldos adeudados más bajos, junto con cuotas de créditos menores, son créditos con mayor antigüedad, por lo que tienen menos cuotas pendientes y tienen mayor amortización. A su vez, tanto a nivel de salario bruto, como neto y líquido, los malos pagadores tienen menores salarios, lo cual es el resultado esperado. Pese a lo anterior, al observar el comportamiento de estos indicadores respecto a deudores pensionados y con condición laboral activa, se obtiene que la tendencia no es la misma entre ambas poblaciones, ya que para deudores pensionados se tienen diferencias más bajas entre

buenos y malos pagadores, y con una relación contraria a la de deudores activos y a la del total de la población, lo que lleva a descartar estas variables como variables explicativas.

Sobre las variables explicativas referentes la morosidad de las operaciones de crédito, se obtiene el resultado esperado, en cuanto que los malos pagadores mantienen niveles de morosidad promedio mucho mayores a los buenos pagadores.

Por otro lado, al aplicar los modelos de aprendizaje supervisado sobre las variables cuantitativas restantes se determina que, por medio del análisis de componentes principales, se tiene la presencia de una sola dimensión con un factor que explica por sí solo hasta 80% de la variancia. Al aplicar el análisis de clúster por medio de la metodología de k medias (ver Cuadro 13), se conforman 3 grupos que diferencian de buena forma según la incidencia de malos pagadores, con niveles de 0,53%, 11,77% y 77,66%, según grupos de baja, media y alta incidencia respectivamente. Dicha conformación fue utilizada como una variable explicativa adicional en los modelos de aprendizaje supervisado, sin embargo, se determinó que no se generaba un aporte adicional para mejorar los resultados de los indicadores de clasificación.

En cuanto al análisis según los modelos de aprendizaje supervisado, se concluye que el modelo bayesiano arroja los mejores resultados de clasificación, con indicadores de clasificación global y de buenos pagadores similares a los restantes modelos, pero con un indicador promedio de clasificación de malos pagadores de 81,95%, siendo claramente superior respecto a los demás modelos.

Aplicando el modelo obtenido sobre la base de datos de validación, correspondiente a las observaciones del año 2023, se generan las probabilidades de incumplimiento predichas, y se generan categorías de riesgo a partir de ellas. Para la generación de las categorías de riesgo se utilizan las probabilidades marcadas por los cuartiles 90,00; 92,50 y 97,50, además de utilizar el corte de 50,00% para definir las dos categorías de mayor riesgo, y en conjunto crear cinco categorías de riesgo. El análisis de las distribuciones de las variables explicativas según las categorías de riesgo construidas admite concluir que dichas categorías permiten medir de manera razonable el nivel de riesgo de las operaciones de crédito.

Por último, al llevar este resultado sobre los datos nuevos, para el proceso de cálculo de pérdida esperada de la cartera, se obtiene que el monto de pérdida esperada recalculado es en promedio ₡5.076 millones más bajo, lo que lleva a reducir el indicador de cobertura (respecto al saldo con mora mayor a 90 días de atraso), de 4,31 veces a 2,90 veces en promedio, lo cual se ubica dentro del nivel de apetito al riesgo definido institucionalmente (al menos 2 veces), por lo que se concluye que el resultado es viable.

La propuesta presentada permite una mejora del esquema de seguimiento y control del nivel de riesgo de crédito, siendo más ajustado a la realidad respecto al modelo actual, proporciona una metodología de calificación de riesgo novedosa con probabilidades de incumplimiento claramente diferenciadas, y favorece para alinear los esfuerzos hacia un mismo resultado, el cual es maximizar el beneficio, ya que permite la liberación de recursos para ser mejor aprovechados en oportunidades de inversión y mejorar el rendimiento de la cartera.

A pesar de lo anterior, dentro de las posibles limitaciones y consideraciones para trabajos futuros, que para esta investigación no pudieron ser abordados por efectos de tiempo, se resalta la oportunidad de profundizar en técnicas para el manejo del desbalance de clases, como métodos de sobre muestreo o submuestreo y usos de una mayor gama de modelos de potenciación que permitan asignar costos a las clases minoritarias. La diferencia entre la población de deudores pensionados con deudores no pensionados plantea la posibilidad de realizar la construcción de un *score* de crédito para las dos poblaciones por separado, o bien, explorar la posibilidad de crear índices según análisis de componentes principales o análisis de clúster que logren aislar esta relación para ser usada en los modelos. Por otra parte, en cuanto al análisis del poder predictivo de las variables, es posible incluir una mayor variedad de técnicas como análisis de correlaciones y dependencia de variables, métricas individuales exploratorias en modelos sin calibrar para analizar el efecto e importancia de las distintas variables, entre otros, estos elementos pueden ser considerados en futuros ejercicios. Además, en cuanto al análisis de clúster, para futuros análisis es posible profundizar en técnicas que combinen variables categóricas y continuas por medio de métricas de distancias como la distancia de *Gower*, que no asume distribuciones normales, o métodos más flexibles

como el Análisis de Clústeres en Dos Fases, que puede manejar mezclas de variables categóricas y continuas.

A modo de recomendación, se resaltan dos elementos, (1) la información analizada respecto de los deudores estudiados, permitió realizar la construcción de un *score* de crédito de seguimiento que permita modelar la probabilidad de incumplimiento del deudor, donde inicialmente se tuvo a disposición hasta 29 variables, sin embargo, resulta relevante contar con información adicional sobre el nivel de endeudamiento de los deudores en el sistema financiero, ya que dichos datos solamente se recolectan al momento del otorgamiento del crédito y no existe un proceso de captura de datos de este tipo para efectos de seguimiento, lo cual podría favorecer para realizar un análisis más exhaustivo sobre la probabilidad de que un deudor no haga frente a sus obligaciones de crédito.

Por otra parte, (2) se comprende que dar un salto de un esquema tradicional, a un esquema como el propuesto, puede abrumar a los tomadores de decisiones en las diferentes instituciones, por lo que, una medida que puede favorecer para generar mayor confianza, es desarrollar el esquema propuesto a modo de plan piloto y de forma paralela al esquema tradicional, para un periodo no menor a un año, con el propósito de generar información y estudiar de mejor manera las ventajas y desventajas de ambos métodos, y que los resultados favorezcan la toma de decisiones basados en la mejor información disponible, lo anterior en un marco que considere lineamientos para minimizar el riesgo del modelo, con seguimientos periódicos y actualizaciones en periodos no mayores a dos años.

BIBLIOGRAFÍA

- Amazon Web Services. (2025). *Amazon*. Obtenido de Guía para desarrolladores: Amazon Machine Learning: https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf#cross-validation
- Andreeva, G. (2006). *The Journal of the Operational Research Society*. Obtenido de European Generic Scoring Models Using Survival Analysis: <https://www.jstor.org/stable/4102368>
- Androutsopoulos, I., Koutsias, J., Chandrinos, K., & Spyropoulos, C. (2008). *An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages*. Obtenido de National Centre for Scientific Research “Demokritos”: <https://arxiv.org/pdf/cs/0008019>
- Angarita Villamizar, J. C., & Marín Guerra, R. A. (2019). *Modelo de predicción de incumplimiento de pago en sector de microfinanzas*. Obtenido de Universidad de los Andes: <https://repositorio.uniandes.edu.co/server/api/core/bitstreams/7f7c9b50-83b6-4cb5-977d-260394dac0cf/content>
- Araf, I., Idri, A., & Chairi, I. (2024). *Cost-sensitive learning for imbalanced medical data: a review*. Obtenido de Artificial Intelligence Review: https://link.springer.com/article/10.1007/s10462-023-10652-8?utm_source=chatgpt.com
- Arafat , H., Hassanat, A., Lasassmeh, O., Tarawneh, A., Bashir, M., Eyal, H., & Prasath, S. (2019). *Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review*. Obtenido de <https://arxiv.org/pdf/1708.04321>
- Baesens, B. (2003). *KATHOLIEKE UNIVERSITEIT LEUVEN*. Obtenido de DEVELOPING INTELLIGENT SYSTEMS FOR CREDIT SCORING USING MACHINE LEARNING TECHNIQUES: <https://www.dataminingapps.com/wp-content/uploads/2015/04/Phd-Bart-Baesens.pdf>
- BBVA. (2025). *Grupo Financiero BBVA*. Obtenido de Gestion del Riesgo: Riesgo de Credito: <https://accionistaseinversores.bbva.com/microsites/bbvain2013/es/G/c2.html#:~:text>

=Existen%20diferentes%20tipos%20de%20scoring,las%20que%20dispone%20el%20Grupo.

Bravo, L., Tovar, R., & Achcar, J. (2015). *Estimación del Valor Predictivo Positivo de la Colangiopancreatografía Magnética utilizando metodos de Bayes*. Obtenido de Universidad del Valle Santiago de Cali: <http://www.scielo.org.co/pdf/rmri/v21n2/v21n2a05.pdf>

CAHUEÑAS IGUAGO, M. I. (Octubre de 2024). *CONSTRUCCIÓN DE UN MODELO SCORING DE SEGUIMIENTO PARA LA CARTERA DE MICROCRÉDITO CON REDES NEURONALES EN LA COOPERATIVA DE AHORRO Y CRÉDITO 23 DE JULIO LTDA*. Obtenido de Facultad de Ciencias de la Escuela Politecnica Nacional: <https://bibdigital.epn.edu.ec/bitstream/15000/26176/1/CD%2014867.pdf>

Carvalho , M., Pinho, A., & Bras, S. (2025). *Resampling approaches to handle class imbalance: a review from a data perspective*. Obtenido de Journal of Big Data: https://link.springer.com/article/10.1186/s40537-025-01119-4?utm_source=chatgpt.com

Chen, T., & Guestrin, C. (2016). *University of Washington*. Obtenido de XGBoost: A Scalable Tree Boosting System: <https://arxiv.org/pdf/1603.02754>

Consejo Nacional de Supervicion del Sistema Financiero. (1 de enero de 2024). *ACUERDO SUGEF 2-10 REGLAMENTO SOBRE ADMINISTRACION INTEGRAL DE RIESGO*. Obtenido de SUGEF: [https://www.sugef.fi.cr/ver/normativa/normativa_vigente/SUGEF%202-10%20\(v29%201%C2%B0%20de%20enero%20de%202024\).pdf](https://www.sugef.fi.cr/ver/normativa/normativa_vigente/SUGEF%202-10%20(v29%201%C2%B0%20de%20enero%20de%202024).pdf)

Consejo Nacional de Supervisión del Sistema Financiero. (2025). *SUGEF*. Obtenido de ACUERDO CONASSIF 14-21 REGLAMENTO SOBRE CÁLCULO DE ESTIMACIONES CREDITICIAS: [https://www.sugef.fi.cr/ver/normativa/normativa_transversal/documentos/CONASSIF%2014-21%20\(v12%2026%20de%20mayo%20de%202025\).pdf](https://www.sugef.fi.cr/ver/normativa/normativa_transversal/documentos/CONASSIF%2014-21%20(v12%2026%20de%20mayo%20de%202025).pdf)

CORTES IZASMENDI, J. C. (octubre de 2018). *CREDIT SCORING: MAQUINAS DE SOPORTE VECTORIAL CONTRA LOS MODELOS CLASICOS DE REGRESION*. Obtenido de Benemérita Universidad Autonoma de Puebla:

<https://repositorioinstitucional.buap.mx/server/api/core/bitstreams/be3ce304-d597-49e2-ab83-7b586654e9eb/content>

COVER, T. (1967). *Nearest Neighbor Pattern Classification* . Obtenido de Stanford University: <https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf>

Data Scientist. (2025). *Data Scientist: Cross-Validation : definición e importancia en Machine Learning*. Obtenido de <https://datascientest.com/es/cross-validation-definicion-e-importancia#>

Escalona, A. (2011). *Uso de los modelos credit scoring en microfinanzas*. Obtenido de Colegio de Postgraduados: http://193.122.196.39:8080/bitstream/handle/10521/414/Escalona_Cortes_A_MC_Estadistica_2011.pdf?sequence=1&isAllowed=y

Etim, R., & Damidele, K. (2022). *Development of Credit Scoring Model for Borrowers Using Machine Learning Techniques*. Obtenido de University of Technology Akure: https://www.researchgate.net/publication/361293278_Development_Of_Credit_Scoring_Model_For_Borrowers_Using_Machine_Learning_Techniques

Federal Reserve, Office of the Comptroller of the Currency. (2011). *Federal Reserve*. Obtenido de <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>

Fix, E., & Hodges, J. (1951). *DISCRIMINATORY ANALYSIS NONPARAMETRIC DISCRIMINATION* . Obtenido de University of California, Berkeley: <https://apps.dtic.mil/sti/pdfs/ADA800276.pdf>

García Ojeda, J. G. (2023). *DETERMINACIÓN DE ALGORITMOS DE CLASIFICACIÓN ÓPTIMOS PARA LA EVALUACIÓN DE RIESGO CREDITICIO*. Obtenido de Universidad Nacional de Cuyo: https://bdigital.uncu.edu.ar/objetos_digitales/19342/determinacin-de-algoritmos-de-clasificacin-ptimos-para-la-evaluacin.pdf

Gomez Betancur, J. C. (2024). *Clasificación de créditos de libranza negociados en el mercado secundario colombiano, aplicando técnicas de aprendizaje supervisado*. Obtenido de UNIVERSIDAD EAFIT: <https://repository.eafit.edu.co/server/api/core/bitstreams/6e7eae21-864e-4676-a545-675c964eff76/content>

- Granja García, V., & García Montealegre, N. (Agosto de 2020). *MODELO DE SEGUIMIENTO DE RIESGO DE CRÉDITO PARA EL CLIENTE INDEPENDIENTE, PARA UNA ENTIDAD FINANCIERA DEL VALLE DEL CAUCA*. Obtenido de Universidad EAFIT Colombia: <https://repository.eafit.edu.co/server/api/core/bitstreams/a26b08ac-04be-450c-95da-85cdf0996714/content>
- Gutiérrez Girault, M. A. (2007). *Banco Central de la República Argentina (BCRA)*. Obtenido de Modelos de Credit Scoring Qué, Cómo, Cuándo y Para Qué: <https://www.bcra.gov.ar/Pdfs/Publicaciones/CreditScoring.pdf>
- Hand, D., & Henley, W. (1997). *Journal of the Royal Statistical Society. Series A (Statistics in Society)*. Obtenido de Statistical Classification Methods in Consumer Credit Scoring: A Review: <https://www.jstor.org/stable/2983268>
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Springer.
- Hernández, O. (2013). *TEMAS DE ANÁLISIS ESTADÍSTICO MULTIVARIANTE*. San Jose: UCR.
- Junta de Pensiones y Jubilaciones del Magisterio Nacional. (2023). *Declaratoria Apetito Riesgo Institucional*. Obtenido de JUPEMA: <https://www.juntadepensiones.cr/admin/archivos/contenido/archivo/743>true>
- Junta de Pensiones y Jubilaciones del Magisterio Nacional. (2024). *Evaluación Actuarial 2024*. Obtenido de JUPEMA: <https://www.juntadepensiones.cr/admin/archivos/contenido/archivo/816>true>
- Junta de Pensiones y Jubilaciones del Magisterio Nacional. (2024). *Reporte de Gestión 2024*. Obtenido de JUPEMA: <https://www.juntadepensiones.cr/admin/archivos/contenido/archivo/821>true>
- Junta de Pensiones y Jubilaciones del Magisterio Nacional. (2025). *JUPEMA*. Obtenido de Informe Azul: <https://www.juntadepensiones.cr/admin/archivos/contenido/archivo/832>true>
- Manning, C., Raghavan, P., & Schütze, H. (2008). *Cambridge University*. Obtenido de Introduction to Information Retrieval: <https://www-nlp.stanford.edu/IR-book/>

- Microsoft. (2024). *Learn Microsoft*. Obtenido de Machine Learning: Modelo de validación cruzada: <https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/cross-validate-model?view=azureml-api-2>
- Olabenjo, B. (2016). *Applying Naïve Bayes Classification to Google Play Apps Categorization*. Obtenido de University of Saskatchewan: <https://arxiv.org/pdf/1608.08574>
- Organización Internacional de Normalización. (2018). *ISO: Normas mundiales para bienes y servicios de confianza*. Obtenido de Gestión de riesgos: <https://www.iso.org/obp/ui/es/#iso:std:iso:31000:ed-2:v1:es>
- Peña Espinoza, L. A. (Octubre de 2013). *Matrices de Transición del Crédito en Nicaragua*. Obtenido de Banco Central de Nicaragua: https://www.bcn.gob.ni/sites/default/files/documentos/DT-31_Matrices_de_Transicion_del_Credito_en_Nicaragua.pdf
- Polonio, M. (2022). *Estudio del Análisis Discriminante. Aplicación a datos reales*. Obtenido de Universidad de Granada: https://digibug.ugr.es/bitstream/handle/10481/76620/TFG_Maria.pdf?sequence=1&isAllowed=y
- Qing-Yan, Y., Jiang-She, Z., Chun-Xia, Z., & Sheng-Cai, L. (2013). *An Empirical Study on the Performance of Cost-Sensitive Boosting Algorithms with Different Levels of Class Imbalance*. Obtenido de Mathematical Problems in Engineering: <https://onlinelibrary.wiley.com/doi/epdf/10.1155/2013/761814>
- Robles, E. (2014). *EL SISTEMA FINANCIERO COSTARRICENSE EN LOS ÚLTIMOS 25 AÑOS: SISTEMA DE PENSIONES*. Obtenido de Academia de Centroamerica : <https://www.academiaca.or.cr/wp-content/uploads/2017/03/sistema-de-pensiones.pdf>
- Robles, E. (2019). *EL SISTEMA DE PENSIONES EN COSTA RICA Y SUS DESAFÍOS*. Obtenido de Academia de Centroamérica: <https://www.academiaca.or.cr/wp-content/uploads/2020/03/El-sistema-de-pensiones-en-Costa-Rica-y-sus-desaf%C3%ADos.pdf>

- Saito, T., & Rehmsmeier, M. (2015). *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. Obtenido de PLOS ONE RESEARCH ARTICLE: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4349800/pdf/pone.0118432.pdf>
- Schneider, K. (2004). *A comparison of event models for naive bayes anti spam email filtering*. Obtenido de University of Passau: <https://dl.acm.org/doi/pdf/10.3115/1067807.1067848>
- Támara Ayús, A. L. (Octubre de 2016). *El método popperiano en la estimación de la probabilidad de incumplimiento de un deudor*. Obtenido de Departamento de Finanzas de la Universidad EAFIT: <https://dialnet.unirioja.es/descarga/articulo/5757291.pdf>
- Tamara, A., Aristizábal, R., & Velásquez, E. (mayo de 2012). *MATRICES DE TRANSICIÓN EN EL ANÁLISIS DEL RIESGO CREDITICIO COMO ELEMENTO FUNDAMENTAL EN EL CÁLCULO DE LA PÉRDIDA ESPERADA EN UNA INSTITUCIÓN FINANCIERA COLOMBIANA*. Obtenido de Revista Ingenierías Universidad de Medellín: <http://www.scielo.org.co/pdf/rium/v11n20/v11n20a09.pdf>
- Toledo, E., & León, V. (2021). *Matriz de transición: una herramienta para determinar la probabilidad de default en las instituciones microfinancieras*. Obtenido de https://download.ssrn.com/22/01/12/ssrn_id4007449_code1327823.pdf?response-content-disposition=inline&X-Amz-Security-Token=IQoJb3JpZ2luX2VjEDgaCXVzLWVhc3QtMSJHMEUCIFIK2xyIuFUrZy3u0mleYkpLdqIQE3MBN3ZXki6PXRAiEA14tfIJhDJHH5oVU9RRqHAtKnHJhHNhRuH4p6BPweu0q
- Valencia, V., & Jorge, Z. (2013). *Calculo de la Probabilidad de Defaul para una Cartera de Creditos Vehiculares*. Obtenido de Escuela Superior Politecnica del Litoral: <https://www.dspace.espol.edu.ec/rest/bitstreams/79076/retrieve>
- Vargas Sánchez, A., & Mostajo Castelú, S. (2014). *Universidad Privada Boliviana*. Obtenido de MEDICIÓN DEL RIESGO CREDITICIO MEDIANTE LA APLICACIÓN DE MÉTODOS BASADOS EN CALIFICACIONES INTERNAS:

http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2518-44312014000200002

Venkat, S., & Kim, Y. (1987). *The Journal of Finance*. Obtenido de Credit Granting: A Comparative Analysis of Classification Procedures: <https://www.jstor.org/stable/2328378>

ANEXOS

ANEXO 1. Pruebas de aporte marginal de variables de mora máxima y grupo de mora.

Cuadro 24. Indicadores de clasificación alternando variables de mora máxima y grupo de mora según métodos de clasificación, enero y diciembre 2023

Métodos	Modelo	Clasificación global	Clasificación Buenos	Clasificación Malos
Arboles de decisión	Completo	98,23%	99,37%	66,67%
	Sin mora máxima	98,23%	99,37%	66,67%
	Sin grupo deudor	98,23%	99,37%	66,67%
	Sin mora máxima y grupo deudor	98,23%	99,37%	66,67%
Bosques aleatorios	Completo	98,22%	99,29%	68,58%
	Sin mora máxima	98,21%	99,28%	68,37%
	Sin grupo deudor	98,23%	99,33%	67,73%
	Sin mora máxima y grupo deudor	98,24%	99,30%	68,79%
AdaBoost	Completo	98,19%	99,25%	68,58%
	Sin mora máxima	98,21%	99,37%	65,82%
	Sin grupo deudor	98,17%	99,27%	67,52%
	Sin mora máxima y grupo deudor	98,22%	99,30%	68,37%
XGBoost	Completo	98,22%	99,32%	67,52%
	Sin mora máxima	98,25%	99,34%	67,73%
	Sin grupo deudor	98,20%	99,32%	67,09%
	Sin mora máxima y grupo deudor	98,19%	99,30%	67,30%
KNN	Completo	98,05%	99,23%	65,39%
	Sin mora máxima	98,09%	99,21%	66,88%
	Sin grupo deudor	98,10%	99,30%	64,97%
	Sin mora máxima y grupo deudor	98,10%	99,21%	67,09%
SVM	Completo	98,25%	99,36%	67,30%
	Sin mora máxima	98,22%	99,35%	66,88%
	Sin grupo deudor	98,25%	99,33%	68,37%
	Sin mora máximo y grupo deudor	98,25%	99,32%	68,37%
Logística	Completo	98,22%	99,40%	65,18%
	Sin mora máxima	98,21%	99,40%	65,18%
	Sin grupo deudor	98,08%	99,39%	61,78%
	Sin mora máxima y grupo deudor	98,08%	99,40%	61,57%
	Completo	95,41%	95,80%	84,71%

Métodos	Modelo	Clasificación global	Clasificación Buenos	Clasificación Malos
	Sin mora máxima	95,80%	96,24%	83,65%
	Sin grupo deudor	95,69%	96,12%	83,65%
	Sin mora máxima y grupo deudor	96,25%	96,73%	83,01%
LDA	Completo	98,16%	99,28%	67,09%
	Sin mora máxima	98,15%	99,27%	67,09%
	Sin grupo deudor	98,06%	98,95%	73,46%
	Sin mora máxima y grupo deudor	98,10%	98,99%	73,46%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

ANEXO 2. Resultados fase de calibración de parámetros con datos de prueba.

Cuadro 25. Indicadores de clasificación alternando parámetros según métodos de clasificación, enero y diciembre 2023

Modelo	Parámetros	Clasificación global	Clasificación Buenos	Clasificación Malos
Arboles de decisión	Mínimo de cada nodo			
	2	98,14%	99,28%	65,20%
	50	98,14%	99,28%	65,20%
	500	98,14%	99,28%	65,20%
Bosques aleatorios	Cantidad de arboles			
	500	98,14%	99,30%	67,67%
	800	98,14%	99,30%	67,87%
	1000	98,16%	99,30%	68,07%
AdaBoost	Booster			
	Real	98,18%	99,35%	65,81%
	Discrete	98,17%	99,31%	66,40%
	Gentle	98,15%	99,25%	67,61%
XGBoost	Booster			
	Dart	98,16%	99,29%	66,77%
	Gbtree	98,16%	99,29%	66,77%
	Linear	97,99%	99,42%	58,26%
KNN	Kernel			
	inv	98,17%	99,25%	68,27%
	optimal	98,17%	99,35%	65,54%
	triangular	98,16%	99,30%	66,34%
	rank	98,17%	99,34%	65,67%
	epanechnikov	98,14%	99,30%	66,10%

Modelo	Parámetros	Clasificación	Clasificación	Clasificación
		global	Buenos	Malos
SVM	gaussian	98,15%	99,31%	65,98%
	cos	98,15%	99,30%	66,16%
	rectangular	98,14%	99,32%	65,37%
	biweight	98,14%	99,29%	66,11%
	triweight	98,14%	99,29%	66,10%
	radial	98,21%	99,38%	65,70%
	linear	98,19%	99,35%	65,93%
	polynomial	98,15%	99,31%	66,00%
	sigmoid	97,74%	99,09%	60,25%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

ANEXO 3. Resultados análisis de probabilidad de corte en modelo ganador con datos de prueba.

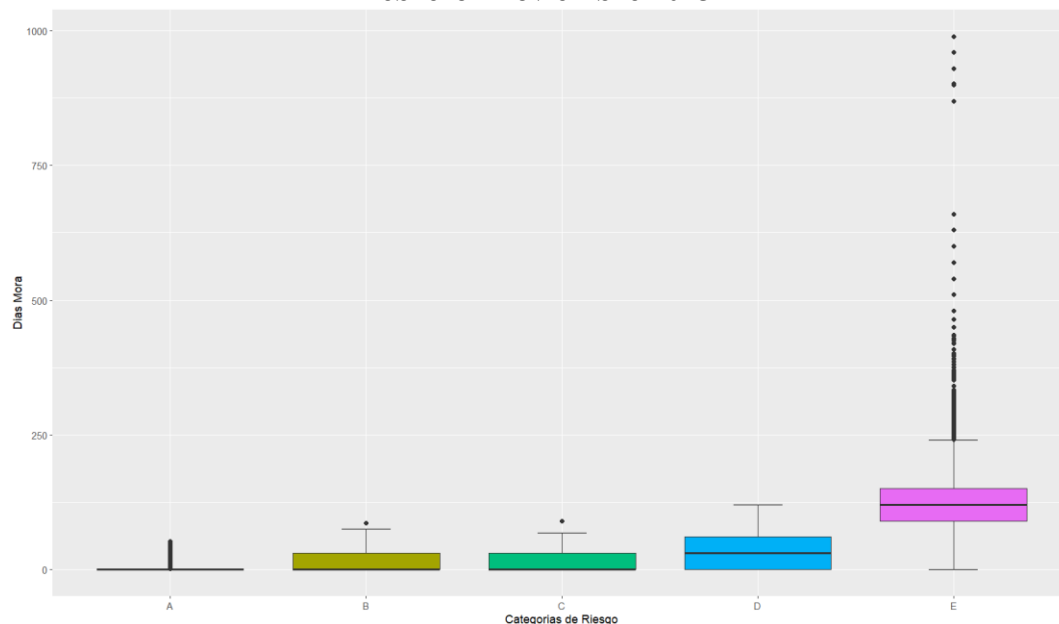
Cuadro 26. Indicadores de clasificación alternando parámetros según métodos de clasificación, enero y diciembre 2023

Corte	Clasificación	Clasificación	Clasificación
	global	Buenos	Malos
1,00	98,04%	99,38%	60,95%
0,95	96,68%	97,30%	79,50%
0,90	96,50%	97,09%	80,01%
0,85	96,42%	97,00%	80,22%
0,80	96,23%	96,79%	80,77%
0,75	96,19%	96,74%	80,77%
0,70	96,10%	96,65%	80,94%
0,65	96,07%	96,61%	80,98%
0,60	96,03%	96,57%	81,03%
0,55	96,00%	96,54%	81,11%
0,50	95,96%	96,49%	81,20%
0,45	95,93%	96,46%	81,28%
0,40	95,87%	96,39%	81,41%
0,35	95,84%	96,36%	81,54%
0,30	95,79%	96,30%	81,71%
0,25	95,65%	96,13%	82,22%
0,20	95,57%	96,04%	82,43%
0,15	95,37%	95,83%	82,77%
0,10	95,24%	95,67%	83,06%
0,05	95,01%	95,43%	83,53%
0,00	3,47%	0,00%	1,00%

Fuente: elaboración propia a partir de datos de crédito de JUPEMA

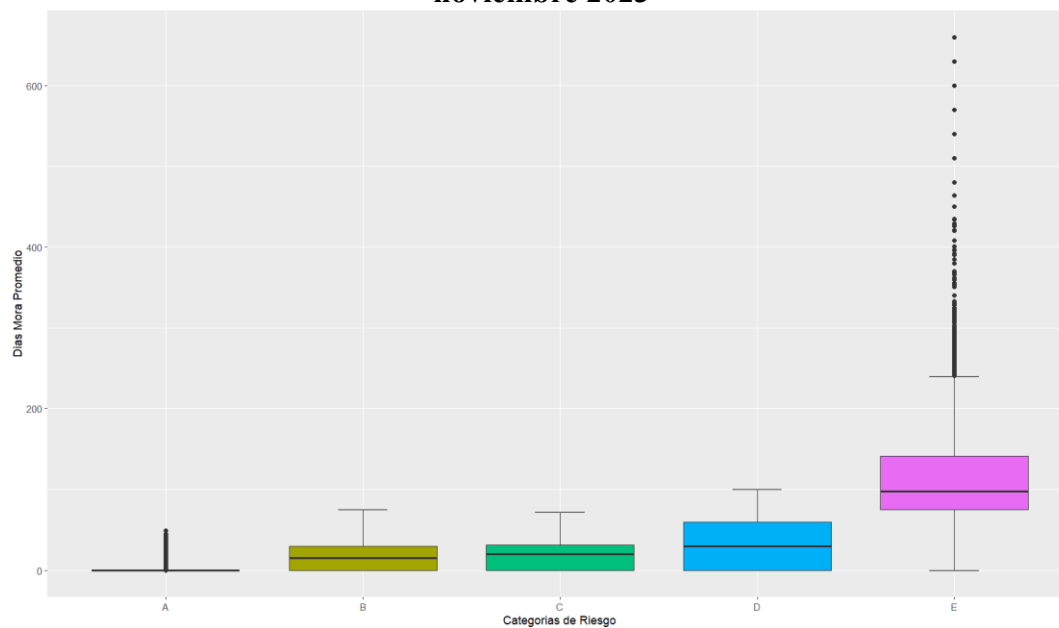
ANEXO 4. Validación de categorías de riesgo según variables del score de crédito para el año 2023.

Gráfico 9.
Distribución de días de mora de operaciones de crédito según categorías de riesgo, febrero - noviembre 2023



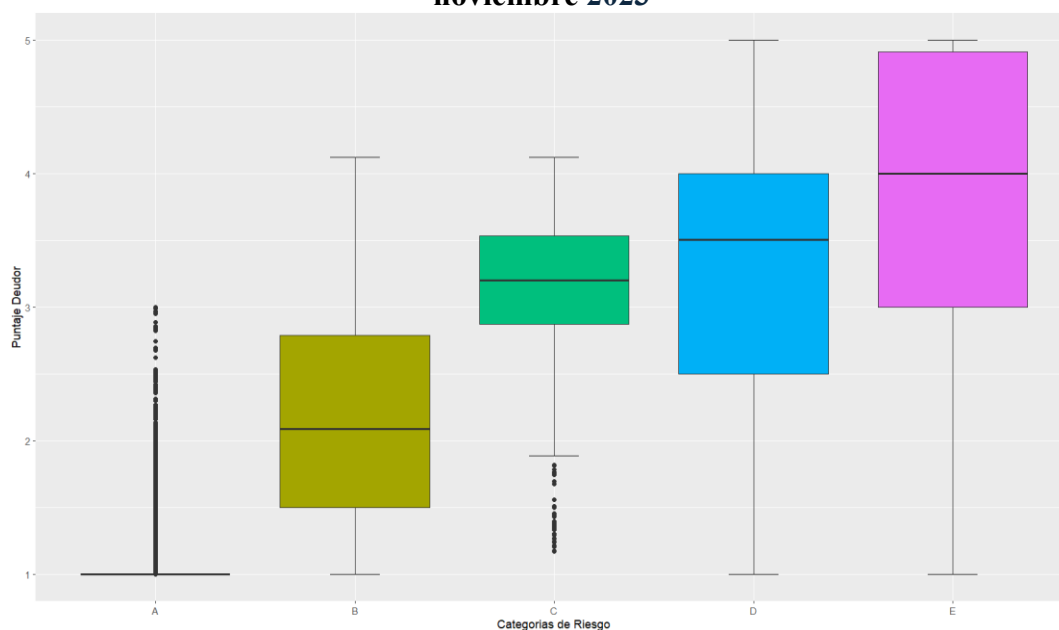
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 10.
Distribución de mora promedio del deudor según categorías de riesgo, febrero - noviembre 2023



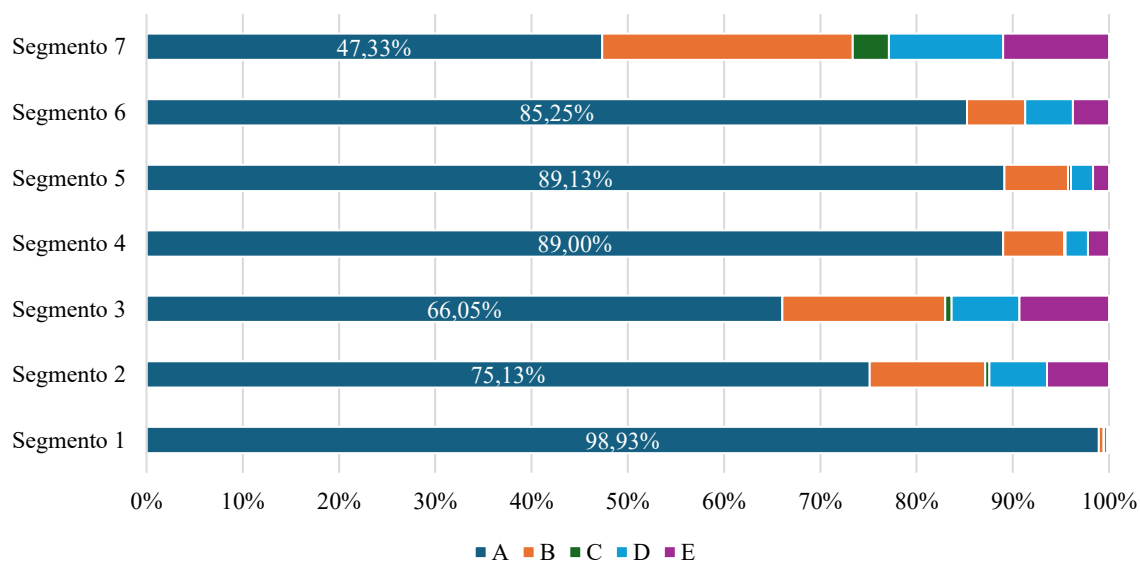
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 11.
Distribución de puntaje histórico deudor según categorías de riesgo, febrero - noviembre 2023



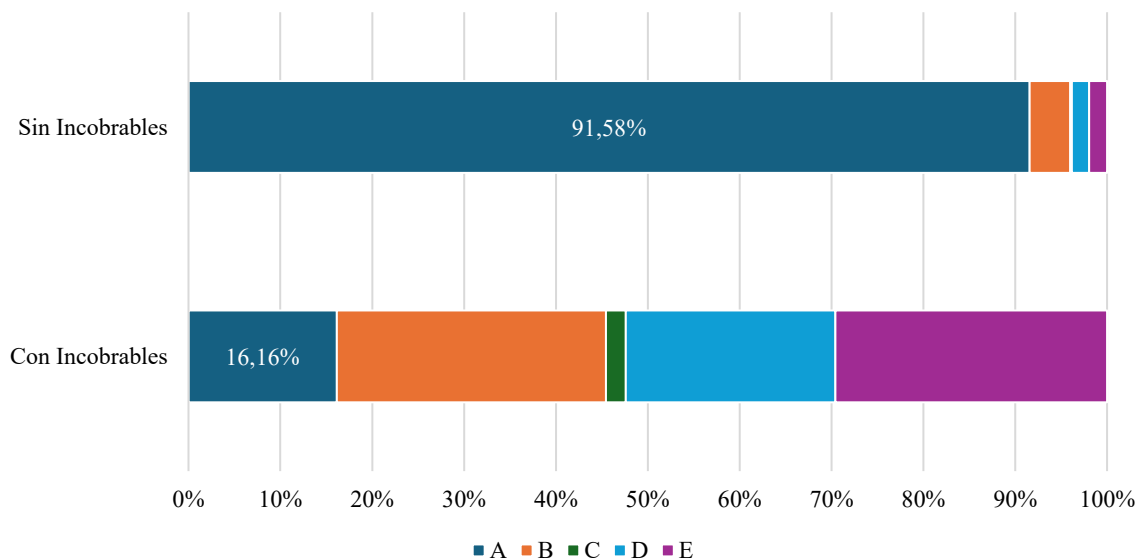
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 12.
Distribución de segmento de mercado según categorías de riesgo, febrero - noviembre 2023



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

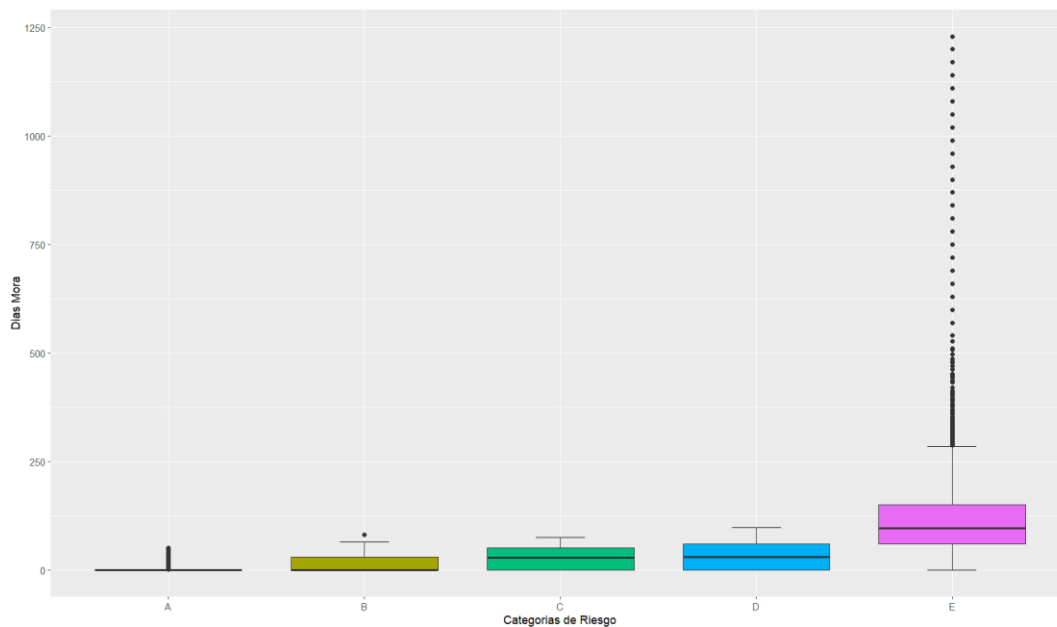
Gráfico 13.
Distribución de tenencia de operaciones incobrables según categorías de riesgo, febrero - noviembre 2023



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

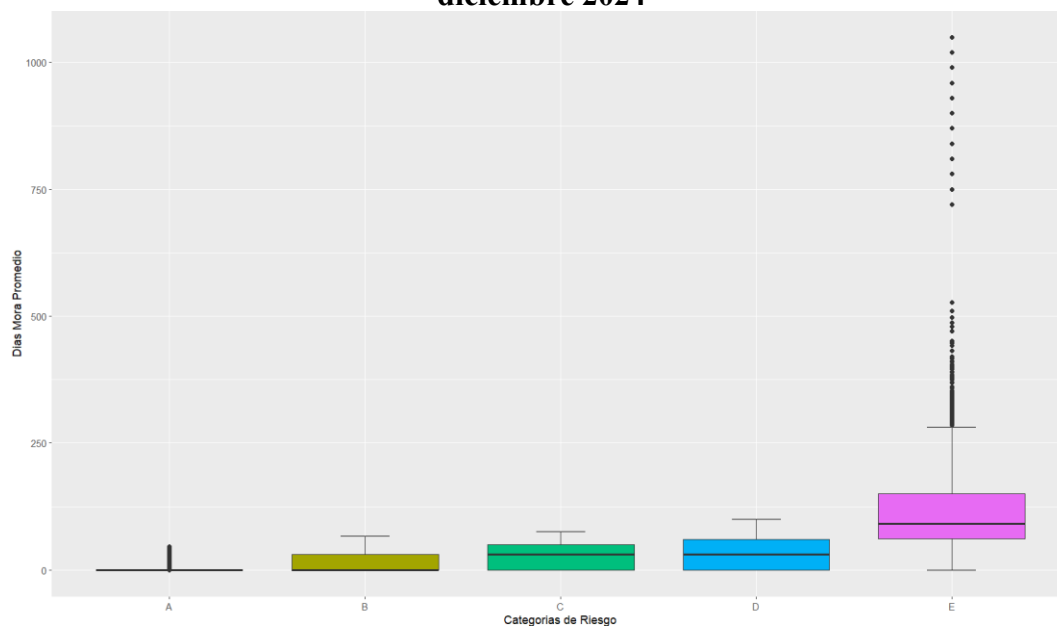
ANEXO 5. Validación de categorías de riesgo según variables del score de crédito para el año 2024.

Gráfico 14.
Distribución de días de mora de operaciones de crédito según categorías de riesgo, enero - diciembre 2024



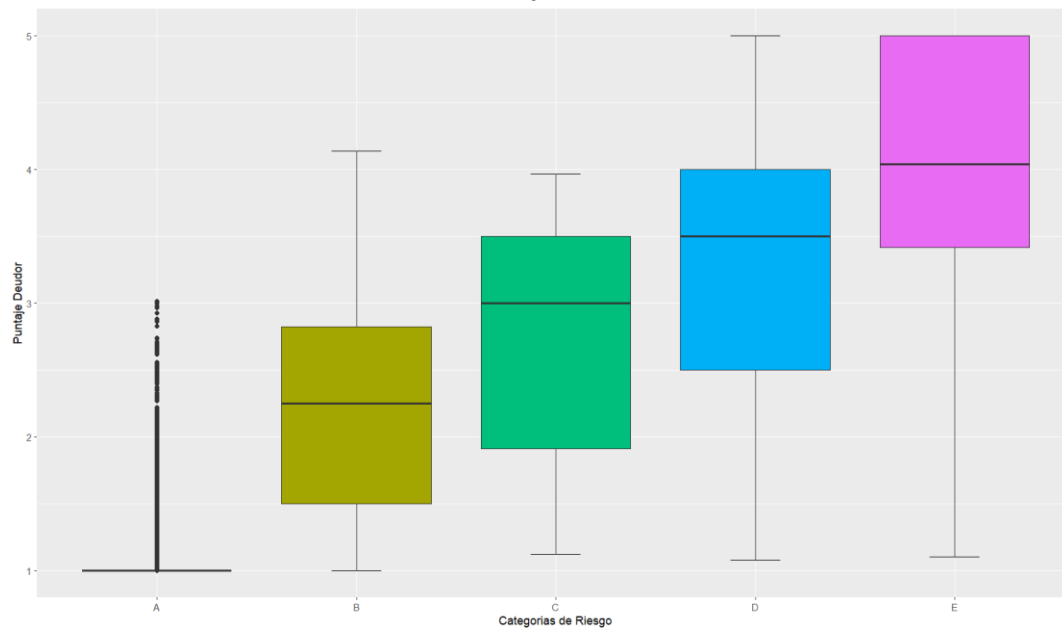
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 15.
Distribución de mora promedio del deudor según categorías de riesgo, enero - diciembre 2024



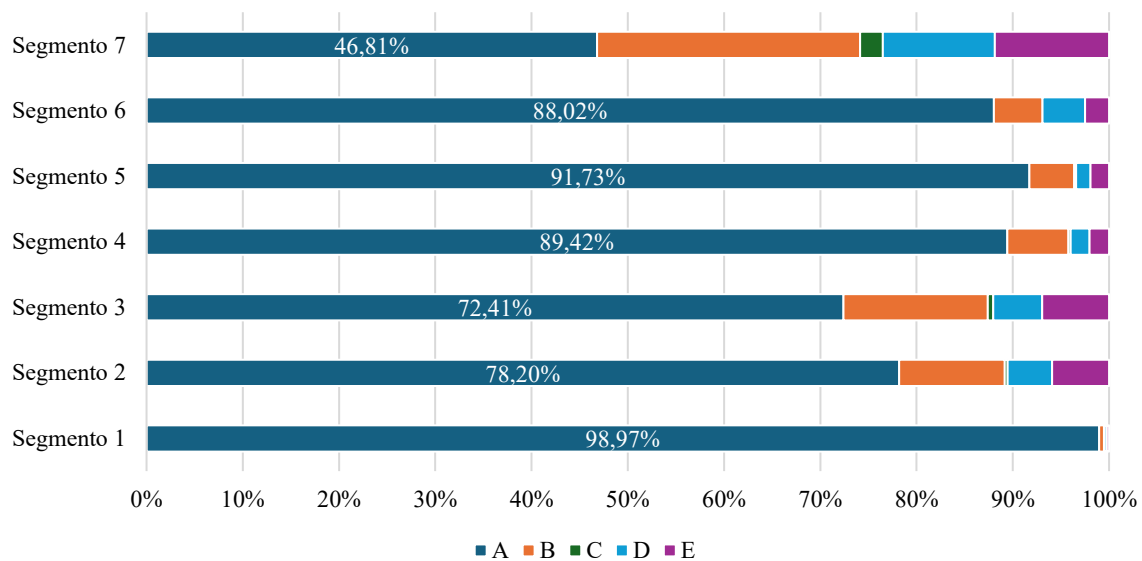
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 16.
Distribución de puntaje histórico deudor según categorías de riesgo, enero - diciembre 2024



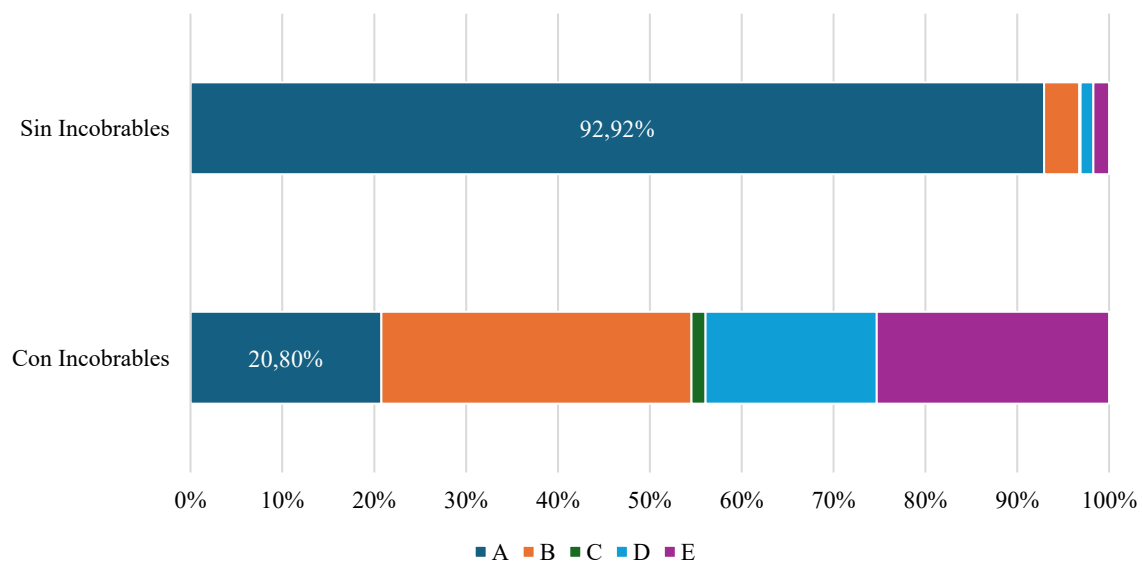
Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 17.
Distribución de segmento de mercado según categorías de riesgo, enero - diciembre 2024



Fuente: elaboración propia a partir de datos de crédito de JUPEMA

Gráfico 18.
Distribución de tenencia de operaciones incobrables según categorías de riesgo, enero - diciembre 2024



Fuente: elaboración propia a partir de datos de crédito de JUPEMA