

Universidad de La Salle

Maestría en Analítica e Inteligencia de Negocios



Título

*Análisis de Variables Determinantes en el Incumplimiento de Pagos de Créditos mediante
Modelos de Clasificación: Caso Rena Ware de Colombia (2022 – 2025)*

Presentado por:

Aura Fabiola Chia Niño

Geraldine Chia Niño

Bajo la tutoría de:

Yeimmy

Katherine Serrano León

Modalidad de grado desarrollo de un proyecto investigativo disciplinar

Bogotá, D.C. 23 de noviembre de 2025

Agradecimientos

Queremos expresar nuestro sincero agradecimiento a nuestra tutora de tesis, la Master. Yeimmy Katherine Serrano León, por su acompañamiento durante este proceso. Su orientación y experiencia en la temática de investigación fueron claves para la realización de este trabajo.

Así mismo, extendemos nuestro agradecimiento al docente Leandro Vivas, quien, amablemente nos ofreció su tiempo, conocimiento, cuestionamientos y aportes que enriquecieron nuestras ideas y dieron un enfoque más sólido al proyecto.

Agradecemos especialmente a nuestros padres, quienes, con su amor profundo, fe y su apoyo incondicional fueron el motor que nos impulsó a culminar este camino y, sus oraciones y palabras de aliento nos reconfortaron en los momentos de dificultad.

A nuestros compañeros de maestría y amigos más cercanos, gracias por acompañarnos en el transcurso de este proceso y por brindarnos las palabras precisas y ánimos para seguir adelante. El camino que emprendimos juntos, con esfuerzo y perseverancia culminó exitosamente.

Finalmente, agradecemos a la empresa Rena Ware de Colombia por abrirnos sus puertas y compartir la información de la compañía necesaria para esta investigación. Su disposición y colaboración fue clave para comprender el negocio y construir una propuesta que aporte valor.

A todos, gracias por acompañarnos durante este proceso y ser parte de este logro.

Dedicatoria

A mi mamá, Myriam, por su amor infinito y su fe silenciosa pero poderosa que me sostuvo aún desde lo intangible y por ser mi mayor apoyo en cada paso de mi vida.

A mi papá, José, por sus consejos sabios y por enseñarme que con esfuerzo, disciplina y tenacidad todo es posible.

A mi hermana, Aura, por ser no solo mi hermana sino también mi compañera en este camino con quien compartí retos, aprendizajes, frustraciones y logros.

Y, en especial, me la dedico a mí misma, por creer, persistir y hacer realidad mi sueño de convertirme en Magíster. Este logro representa mi crecimiento profesional, pero, sobre todo, mi convicción de que puedo alcanzar toda meta que me proponga.

Geraldine Chia Niño

A mis padres, José Nevardo y Miryam por ser mi motor y mi mayor fuente de inspiración. Su amor, esfuerzo y confianza me han acompañado en esta etapa académica y a lo largo de mi formación, recordándome siempre el valor del trabajo y la perseverancia. Gracias por sus enseñanzas y por la seguridad que me brindan de que todo se puede lograr con, disciplina y esfuerzo.

A mi hermana Geraldine, quien no solo ha sido mi soporte incondicional, sino también mi compañera en este proceso académico. Su compañía, amor y paciencia me han sostenido en los momentos difíciles y me han impulsado a seguir adelante con determinación y alegría incluso cuando el camino se tornaba desafiante.

Finalmente, me la dedico a mí misma, por haber creído en mis capacidades, por no rendirme ante los desafíos y por confiar que todo esfuerzo tiene su recompensa. Esta tesis es también un reflejo de mi fortaleza, de mi crecimiento personal y de la certeza de que logro todo lo que me proponga cuando se trabaja con pasión y constancia.

Aura F. Chia Niño

Declaración de originalidad y autonomía

Firma de los autores

Aura Fabiola Chia Niño

Geraldina Chia Niño

Firmado, en la versión PDF, a los 23 días del mes de noviembre del año 2025

Declaración de exoneración de responsabilidad

Aura Fabiola Chia Niño

Geraldine Chia Niño

Firmado, en la versión PDF, a los 23 días del mes de noviembre del año 2025

Tabla de contenido

1. Capítulo I. Contexto	12
a. Planteamiento del Problema.....	12
b. Formulación del problema	14
c. Justificación.....	14
d. Objetivos	16
2. Capítulo II. Fundamentación teórica.....	16
a. Estado del arte	16
b. Marco teórico	24
3. Capítulo III. Diseño metodológico.....	31
a. Enfoque y alcance de la investigación	31
b. Datos de la investigación.....	33
c. Técnicas y modelos	41
4. Capítulo IV. Resultados	48
a. Caracterización de la cartera de créditos.....	48
b. Relación entre variables independientes	54
c. Implementación y evaluación de desempeño de los modelos de clasificación Regresión Logística y Random Forest.....	57
5. Capítulo V. Discusiones	64
6. Capítulo VI. Conclusiones	69

Lista de tablas

Tabla 1. Porcentaje promedio de cartera vencida por altura para los años 2022-2024	13
Tabla 2. Categorías de riesgo de crédito de consumo según los meses de mora.....	26
Tabla 3. Distribución de valores nulos del dataset inicial por variable de modelamiento	34
Tabla 4. Composición de la base de datos antes y después de la depuración.....	35
Tabla 5. Diccionario de datos de base de datos inicial proporcionada por la empresa.	36
Tabla 6. Variables finales utilizadas en el dataset para modelamiento.	38
Tabla 7. Comparación de desempeño de los modelos Random Forest antes y después de aplicado el método de la rejilla	61

Lista de graficas

Ilustración 1. Distribución de los clientes por edad.....	48
Ilustración 2. Distribución de los clientes por género.....	49
Ilustración 3. Distribución de los clientes por estrato.....	50
Ilustración 4. Distribución de los clientes por saldo a financiar y % de cuota inicial	51
Ilustración 5. Distribución de clientes por ingresos totales y número de cuotas del crédito ...	51
Ilustración 6. Distribución de clientes por puntaje crediticio del buro	52
Ilustración 7. Distribución de clientes por incumplimiento.....	53
Ilustración 8. Distribución de clientes por departamento de ubicación e incumplimiento.	54
Ilustración 9. Mapa de correlación entre variables numéricas.....	56
Ilustración 10. Ranking de variables más importantes sobre el incumplimiento	56
Ilustración 11. Métricas generales de los modelos de clasificación con técnicas de balanceo: sin balanceo, class_weight, rus, ros, Smote	57

Ilustración 12. Métricas de los 3 modelos de clasificación con mejor desempeño global:

Random Forest rus, ros y Smote58

Ilustración 13. Curvas ROC de los 3 modelos con mejor desempeño: Random Forest rus, ros y Smote.....59

Ilustración 14. Matrices de confusión de los 3 modelos con mejor desempeño: Random Forest rus, ros y Smote.....59

Ilustración 15. Métricas e hiperparámetros del modelo Random Forest RUS con optimización62

Ilustración 16. Métricas e hiperparámetros del modelo Random Forest ROS con optimización.63

Ilustración 17. Métricas e hiperparámetros del modelo Random Forest Smote con optimización.63

Resumen

La investigación realizada tuvo como finalidad presentar un análisis acerca de las variables que afectan en el incumplimiento de pagos de los créditos que otorga la compañía Rena Ware de Colombia S.A.S. usando modelos de clasificación para predecir la probabilidad de riesgo crediticio de los clientes en el periodo 2022 a junio 2025. Dentro del análisis realizado, se buscó determinar los factores más relevantes en el comportamiento de pago (demográficos y financieros) con el fin de que la empresa pueda fortalecer su gestión de cartera y mejorar los procesos de evaluación de créditos.

El enfoque sobre el cual se desarrolló esta investigación fue un enfoque cuantitativo, con alcance explicativo y predictivo. La base de datos sobre la cual se realizó el estudio corresponde a un dataset suministrado por Rena Ware de 11.860 créditos otorgados durante los años 2022 a junio 2025, tomando cada crédito como una unidad individual de estudio. Dicha base fue constituida una vez realizado un proceso de depuración y normalización de la base de datos con el fin de dejar únicamente las variables más relevantes para el estudio e información completa y consistente. Posteriormente, se aplicaron los modelos de Regresión Logística y Random Forest sin balanceo y con diferentes técnicas de balanceo: Class weight (ponderación de clases), RUS (Random Undersampling), ROS (Random Oversampling) y Smote (Synthetic Minority Oversampling Technique) y se evaluó su desempeño mediante métricas como AUC, F1-score, precisión y sensibilidad.

Los resultados evidenciaron que ambos modelos con diferentes técnicas de balanceo lograron un desempeño aceptable con valores AUC entre 0.76 y 0.84. Sin embargo, el modelo Random Forest con técnicas de balanceo ROS, RUS y Smote obtuvo el mejor desempeño por su equilibrio entre precisión y sensibilidad (AUC promedio de 0.835, Recall de 0.628, Accuracy de 0.777 y F-1 score promedio de 0.632). Aunque finalmente, para el contexto de Rena Ware, el

modelo que mejor se adapta es el Random Forest con balanceo Smote cuyo AUC fue 0.829 por su capacidad de detección de morosos sin ser radical y adverso al riesgo. De igual forma, se implementó el método de la rejilla en busca de optimización de los modelos, no obstante, los resultados obtenidos no fueron significativamente mejores en comparación a los modelos iniciales sin optimización.

Así mismo, el modelado permitió identificar las variables con mayor incidencia en el incumplimiento encontrando que, el puntaje crediticio, la cuota inicial, la tasa efectiva anual, el saldo a financiar, los ingresos y la edad son las variables con mayor peso (0.78 en conjunto) y las más determinantes a la hora de evaluar el incumplimiento. En caso de que la empresa Rena Ware desee implementar dentro de sus procesos de evaluación crediticio algún modelo de clasificación, se recomienda el uso del Random Forest ya que, le permitirá clasificar y priorizar los casos de mayor riesgo.

Finalmente, es importante mencionar que existen limitaciones con el estudio realizado relacionados con la disponibilidad y alcance de los datos ya que, la información utilizada proviene exclusivamente de los registros internos de la empresa en el momento de la aprobación de los créditos y no incorporan variables macroeconómicas o externas que pudiesen robustecer los modelos. Futuras investigaciones podrían ampliar la base de datos e integrar esta información para obtener un modelo más preciso y mejorar así los resultados.

Palabras clave:

Modelos de clasificación, Incumplimiento de pagos, Riesgo de Crédito, Rena Ware, Random Forest, Regresión Logística.

Introducción

En Colombia, las compañías que otorgan crédito de manera directa a sus clientes enfrentan el desafío de gestionar muy de cerca su cartera y minimizar las pérdidas. Hoy en día, en donde las empresas buscan tener un alto desempeño y dependen de la información para poder tomar decisiones eficaces, la capacidad de anticipar riesgos financieros es un factor clave para que las empresas sean estables y sostenibles.

Es así, como Rena Ware de Colombia, una compañía que se ha dedicado a la fabricación y posterior venta de utensilios de cocina premium y que se encuentra registrada como una empresa del mercado multinivel, ha desarrollado su modelo de negocio por más de 30 años en un modelo comercial basado en la financiación de sus productos. No obstante, el aumento en su cartera vencida ha dejado ver la necesidad de incorporar herramientas analíticas que fortalezcan la evaluación de los clientes en cuanto al riesgo crediticio y faciliten así la toma de decisiones.

En ese sentido, el incumplimiento de pagos es un tema de relevancia tanto en la parte teórica como práctica. Desde la perspectiva científica, existen teorías económicas como la información asimétrica (Akerlof, 1970), la selección adversa y el riesgo moral (Stiglitz & Weiss, 1981), la teoría del agente y el principal (Jensen et al., 1976) y la teoría de segmentación de mercados (Smith 1956) que permiten comprender conceptualmente el fenómeno del riesgo crediticio y sus efectos. Es importante resaltar que el presente estudio no tiene como propósito comprobar o validar dichas teorías, sino evaluar la incidencia de variables demográficas y crediticias en el incumplimiento de pagos de los créditos a través de la comparación de modelos predictivos aplicado al análisis del incumplimiento de pagos de los clientes de Rena Ware. Por esta razón, las teorías revisadas se utilizan únicamente como un marco de referencia para comprender el fenómeno del riesgo crediticio y contextualizar el análisis, sin que los resultados

obtenidos pretendan confirmar o refutar dichos enfoques teóricos. Es así, como aplicar modelos de clasificación permite ofrecer soluciones para predecir la probabilidad de incumplimiento y ayuda a segmentar los clientes según su nivel de riesgo. Esto, con el fin de brindar herramientas a las empresas que les permitan reducir pérdidas y diseñar estrategias efectivas.

Ahora bien, aunque el sector financiero en Colombia ha avanzado a lo largo del tiempo en la implementación de modelos de score crediticio, aún existe una limitante en la aplicación de estas metodologías en empresas como Rena Ware que, son compañías no financieras que otorgan crédito directo a sus clientes y que en muchos casos no cuentan con historial crediticio. Esta característica hace que la compañía sea ideal para aplicar modelos de aprendizaje automático pues pueden apoyar a la gestión de la cartera y mitigar el riesgo.

En cuanto al aspecto metodológico, el estudio se enmarca en un enfoque cuantitativo con alcance explicativo y predictivo, haciendo uso de datos reales suministrados por la empresa extraídos del sistema CRM. Dichos datos fueron depurados y balanceados mediante técnicas de balanceo Class weight, Random Oversampling (ROS), Random Undersampling (RUS) y Smote. Posterior a esto, se aplicaron los modelos de clasificación Regresión Logística y Random Forest y se evaluó su desempeño a través de métricas como precisión, sensibilidad, F1-score y área bajo la curva ROC (AUC).

Una vez aplicado lo anterior, se obtuvo como resultado que el modelo con mejor desempeño fue el Random Forest cuyas métricas AUC en promedio de 0.835 con las diferentes técnicas de balanceo (Class_weight, RUS, ROS y Smote). Así mismo, de las variables analizadas se determina que el puntaje crediticio, cuota inicial, la tasa efectiva anual, saldo a financiar, ingresos y edad son las variables de mayor incidencia en el incumplimiento.

Finalmente, el presente trabajo busca evidenciar que la aplicación de herramientas analíticas y modelos de clasificación permite a las empresas fortalecer su gestión del riesgo crediticio. Así mismo, brindar a la compañía Rena Ware una herramienta que le permita mejorar su sostenibilidad financiera y apoyar con datos la toma de decisiones en las áreas de crédito y cartera.

1. Capítulo I. Contexto

a. Planteamiento del Problema

Rena Ware de Colombia S.A.S es una compañía multinacional de venta directa la cual está orientada a la fabricación y venta de utensilios de cocina premium y para el hogar, quien además de ofrecer productos de muy alta calidad, ofrece una oportunidad de emprendimiento para quienes buscan alcanzar el éxito e independencia convirtiéndose en representantes independientes de venta Rena Ware (Rena Ware International INC., 2022). La compañía también lleva más de 80 años en el mercado internacional y 30 años desde su apertura en el mercado colombiano, así como con más de 200.000 clientes a nivel nacional y cerca de 4.000 vendedores independientes Rena Ware.

En cuanto al comportamiento general del sector de venta directa en Colombia al cual pertenece Rena Ware, de acuerdo con cifras de la Asociación Colombiana de Venta Directa Acovedi (2022) y Forbes (2023), en Colombia creció un 14.5% en 2021 y un 8% en el año 2022, reportando más de 6 billones de pesos en ventas. Así mismo, según el informe de la asociación, las mujeres que realizan esta actividad representan el 84% del mercado.

La cartera de consumo en Colombia ha registrado varios cambios en los últimos años influenciados por factores económicos y sociales. En agosto de 2023, se registró un aumento del

27,57% en la cartera vencida en comparación con el año 2022. La cartera de consumo fue la más afectada, con un crecimiento del 46,37%.

Para diciembre de 2024, la Superfinanciera financiera de Colombia registró \$708,8 billones en el total de la cartera bruta, de la cual, el 51,03% corresponde a la modalidad comercial, el 28,52% corresponde a crédito de consumo, el 17,61% a crédito de vivienda y el 2,84% al sector de microcrédito (Superintendencia Financiera de Colombia, 2024).

Con respecto a la cartera de Rena Ware, al cierre de 2024, la empresa reportó un total de cartera cercano a los 20.000 millones de pesos, correspondiente a 10.379 créditos financiados hasta esa fecha. De esta cartera, el 87,51% corresponde a créditos con menos de 30 días de mora o a créditos no vencidos. Así mismo, durante los años 2022 y 2023, el monto promedio de la cartera era aproximadamente de 23.000 millones de pesos, lo que equivale a una disminución del 13,04% en 2024, equivalente a 3.000 millones de pesos.

En relación con la cartera promedio al cierre de los años 2022, 2023 y 2024, según el plazo de vencimiento, se tiene:

Tabla 1. Porcentaje promedio de cartera vencida por altura para los años 2022-2024

Plazo de Vencimiento	Porcentaje de la Cartera Promedio
+30 días	4.77%
+60 días	3.43%
+90 días	2.61%
+120 días	2.19%
+150 días	2.15%
+180 días	1.88%

Fuente: Elaboración propia a partir de datos proporcionados por Rena Ware.

Lo anterior refleja que, en términos monetarios, la cartera vencida representa un promedio aproximado anual de 3.700.000 millones de pesos, lo cual impacta negativamente en la utilidad de la compañía. Por lo tanto, es crucial trabajar en la reducción de este indicador y a su vez, identificar los factores que contribuyen a su aumento con el fin de implementar medidas preventivas y correctivas.

b. Formulación del problema

Según la descripción del problema mencionado previamente surge el siguiente interrogante, el cual corresponde a la formulación del problema:

¿Cuáles son las variables determinantes en el incumplimiento de pagos de créditos de la empresa Rena Ware de Colombia, según los resultados obtenidos mediante modelos de clasificación aplicados a datos del periodo 2022–2025?

c. Justificación

La recuperación de la cartera de crédito es un desafío para las empresas que trabajan con financiamiento a corto y mediano plazo, especialmente aquellas dedicadas a la venta directa, como Rena Ware de Colombia S.A.S. Esta empresa, la cual tiene una trayectoria de más de 30 años en el mercado en la venta de utensilios de cocina premium, obtiene gran parte de sus utilidades operacionales de la recuperación de cartera de sus ventas a crédito. En los últimos años, Rena Ware ha experimentado un crecimiento en la adquisición de nuevos clientes, con un incremento aproximado del 12% anual entre 2022 y 2024. Sin embargo, el 56% de sus ventas corresponden a ventas a crédito, lo que hace que la gestión de la cartera sea una prioridad para la compañía considerando que el aumento de la morosidad en los créditos genera una situación de riesgo en los ingresos y la sostenibilidad de la compañía.

Según los informes más recientes, la cartera vencida de Rena Ware ha aumentado de forma continua afectando negativamente las finanzas y rentabilidad. A finales de 2024, la cartera vencida representaba alrededor de 3.700 millones de pesos, lo que nos demuestra la necesidad de generar de estrategias eficaces para predecir y gestionar los incumplimientos de pago de los clientes y los factores de riesgo que pueden estar afectando la recuperación del crédito. Además, el contexto económico y los cambios que ha tenido la cartera de consumo en Colombia también refuerzan la importancia de reducir el riesgo de crédito y estudiar muy bien al cliente en el momento de aprobar las ventas a crédito.

Para enfrentar este reto, el uso de modelos de clasificación basados en grandes volúmenes de datos, como los de la base de datos de Rena Ware, se presenta como una solución efectiva ya que, a través de estos se logran detectar patrones que se encuentran ocultos en los datos históricos permitiendo predecir la probabilidad de que un cliente incumpla o no sus pagos. Esto se puede lograr analizando las características de aquellos clientes que han incumplido sus pagos, así como las de aquellos que mantienen un buen comportamiento de pago, clasificando a los nuevos clientes en estas categorías según su probabilidad de comportamiento.

El objetivo que se plantea a lo largo de esta investigación es lograr identificar cuáles son las variables determinantes en el incumplimiento de los pagos, como las características demográficas de los clientes y/o crediticias y cuál es el mejor modelo para predecir el incumplimiento de los créditos de Rena Ware, haciendo uso de los datos históricos de la compañía de los años 2022 a 2025.

d. Objetivos

General

Evaluar la incidencia de variables independientes de carácter demográfico y crediticio en el incumplimiento de pagos de créditos otorgados por la empresa Rena Ware de Colombia durante el período 2022–2025, mediante la implementación y comparación de modelos de clasificación.

Específicos

- Caracterizar el comportamiento de la cartera de créditos de los clientes de Rena Ware de Colombia en el período 2022–2025 mediante la aplicación de un análisis exploratorio y descriptivo de los datos.
- Relacionar variables demográficas y crediticias con el incumplimiento de pagos de los clientes de Rena Ware en el periodo 2022-2025.
- Comparar el desempeño de los modelos de clasificación en la predicción del incumplimiento de pagos a través de métricas de ajuste de los modelos.

2. Capítulo II. Fundamentación teórica

a. Estado del arte

El no pago de los créditos de consumo se ha convertido en una problemática tanto para las empresas que ofrecen financiamiento como para la economía a nivel general. El poder anticipar cuando un cliente podría dejar de pagar se ha convertido en un factor estratégico ya que permite a las empresas tomar mejores decisiones, reducir pérdidas y mejorar la gestión de su cartera de créditos.

No obstante, aunque existen muchos estudios sobre el uso de modelos para la revisión y evaluación de los créditos como la Regresión logística, árboles de decisión o los algoritmos de

aprendizaje automático, en su gran mayoría, estos estudios se enfocan en instituciones financieras. Es así, que son pocas las investigaciones que analizan su aplicación en empresas comerciales en Colombia que otorgan crédito directo al consumidor y tienen sus propias bases de datos históricas. Este vacío, es especialmente importante en organizaciones como Rena Ware de Colombia, donde la gestión del crédito se basa en criterios operativos internos, conocimiento del cliente por parte del vendedor independiente y perfiles de clientes que, en algunos escenarios, no están bancarizados.

En esa misma línea, investigaciones relacionadas al riesgo que se tiene al otorgar créditos, han identificado factores que influyen en la probabilidad de que un cliente incumpla un crédito, estos pueden agruparse en tres categorías principalmente: 1) factores sociodemográficos relacionados con la edad, el nivel educativo y el estado civil 2) factores financieros, como el historial crediticio, el nivel de endeudamiento y la capacidad de pago y, 3) factores macroeconómicos, como la tasa de inflación, el desempleo y las condiciones generales la economía.

Dado lo anterior y a partir de estos hallazgos, la revisión del estado del arte busca analizar los modelos de clasificación más aplicados en la predicción del riesgo de crédito, especialmente aquellos que incorporan variables sociodemográficas, financieras y macroeconómicas y que han mostrado buenos resultados al anticipar el incumplimiento. Esto, buscando que las empresas fortalezcan los mecanismos para mitigar riesgos y mejorar sus indicadores. Para organizar la revisión, se plantean cuatro líneas principales: (i) investigaciones que aplican modelos clásicos y modernos para predecir el incumplimiento, (ii) estudios que comparan el desempeño de diferentes algoritmos de clasificación, (iii) experiencias aplicadas a empresas o instituciones del contexto colombiano y (iv) propuestas metodológicas que integran modelos híbridos, técnicas explicables o información alternativa.

Investigaciones que aplican modelos clásicos y modernos para predecir el incumplimiento

En cuanto a la aplicación de modelos clásicos y modernos, Baesens et al. (2003) realizaron uno de los estudios y evaluaciones comparativas de modelos de clasificación más completas aplicados al análisis de riesgo crediticio. En su investigación, analizaron 17 modelos distintos con ocho conjuntos de datos reales de instituciones financieras europeas. Los algoritmos que fueron probados en esta investigación relacionan desde métodos clásicos como la regresión logística y el análisis discriminante, hasta modelos más avanzados como los árboles de decisión, las redes neuronales y las máquinas de soporte vectorial.

Los resultados mostraron que los modelos más complejos, en particular las redes neuronales y las máquinas de soporte vectorial (SVM), lograron las mejores tasas de clasificación y capacidad de discriminación (medida por el AUC). No obstante, los métodos tradicionales como la regresión logística también lograron un gran desempeño, lo que evidencia que no siempre se requieren modelos muy sofisticados o robustos para obtener buenos resultados en el scoring de crédito. Este hallazgo es especialmente relevante para nuestra investigación ya que en el contexto específico de Rena Ware, tanto los modelos clásicos como modernos pueden lograr grandes beneficios financieros si se mejora la predicción.

En una línea similar, Chaabita et al. (2024) hicieron una comparación de tres modelos para la predicción de riesgo crediticio: La regresión logística, las redes neuronales y los árboles de decisión. Su estudio se basó, en evaluar el desempeño de la regresión logística frente a los otros dos modelos basados en inteligencia artificial, para esto, usaron una base de datos con casi 10 mil solicitudes de crédito reales provenientes de un banco en Marruecos. Los resultados obtenidos mostraron una ventaja de los modelos basados en inteligencia artificial donde la red neuronal alcanzó un AUC de 95% mientras que el árbol de decisión alcanzó una precisión del

89,4%. No obstante, aunque la regresión logística no logró igualar estos resultados, los autores destacan su valor por ser un modelo fácil de interpretar y útil en contextos donde la transparencia en la toma de decisiones es prioritaria. En ese sentido, este estudio es muy importante para nuestro proyecto, pues confirma la efectividad de los modelos avanzados en la predicción del incumplimiento de pagos y a su vez, destaca la necesidad de encontrar modelos que permitan tener un equilibrio entre precisión e interpretabilidad de la información.

Por su parte, Coşer et al. (2019) desarrollaron varios modelos de clasificación para predecir el riesgo de incumplimiento de crédito utilizando una base de datos de Lending Club. En su revisión, validaron cuatro algoritmos principales (Regresión Logística, LightGBM, Random Forest y XGBoost) y fueron aplicados bajo tres escenarios de muestreo, dada la alta desproporción entre clientes cumplidos y en mora. El modelo con mejor desempeño fue Random Forest con un AUC de 0.89 en el escenario de muestreo combinado. Más allá de los resultados del modelo, el estudio se destaca por su enfoque y manejo en el desbalance de clases, la validación cruzada del modelo y la selección de variables relevantes. Dado lo anterior, este estudio se convierte en un referente metodológico importantes para proyectos como el de Rena Ware, que enfrenta retos similares al diseñar sistemas de evaluación de créditos sobre datos reales y con distribución desigual entre categorías.

A modo de resumen, los estudios revisados evidencian que tanto los modelos tradicionales como los modelos avanzados de clasificación, tienen fortalezas en la predicción del incumplimiento crediticio. La selección del modelo más adecuado va a depender de las variables y del equilibrio en sus métricas de desempeño tanto en su capacidad predictiva como sensibilidad y exactitud.

Comparación del rendimiento entre diferentes algoritmos de clasificación

En la misma línea del análisis comparativo, Wu & Pan (2021) proponen un modelo integral para la construcción de modelos de scoring crediticio basados en aprendizaje automático, aplicados sobre datos reales de préstamos. Dentro de su estudio, se destaca un proceso de limpieza de datos y también una reducción de variables clave. El estudio comparó los modelos de Regresión Logística, Random Forest y Máquinas de Soporte Vectorial, concluyendo que la regresión logística era el modelo que ofrecía mejor combinación entre rendimiento y estabilidad.

El uso y transformación del modelo estadístico hacia una tarjeta de puntuación crediticia (*scorecard*), fue uno de los más destacados aportes de la investigación, lo que permite aplicar el modelo en la toma de decisiones operativas y combinan ciencia de datos con necesidades en decisiones de riesgo.

Por otro lado, Echeverri Valdes (2006) realizó un análisis comparativo de tres modelos estadísticos para medir el riesgo de incumplimiento crediticio en una entidad financiera del Eje Cafetero. Utilizó información real de clientes con créditos activos aplicando modelos de Regresión Lineal, Análisis discriminante y Modelo Logit. De los modelos validados, el Logit obtuvo los mejores resultados al predecir correctamente el 95.1% de los casos e integrar variables cuantitativas y cualitativas como los días en mora, porcentaje de incumplimiento, número de pagos anuales y valor del crédito. Aunque el estudio corresponde a una entidad financiera, sus hallazgos son valiosos para esta investigación ya que, valida la utilidad del modelo de Regresión Logística en la predicción del comportamiento crediticio en entornos colombianos y demuestra cómo variables operativas como las disponibles en Rena Ware, pueden ser relevantes para anticipar el incumplimiento.

Finalmente, dentro de este enfoque Li & Wu (2024) analizaron el desempeño nueve modelos de aprendizaje automático para predecir el incumplimiento de préstamos, haciendo uso

de una base de datos con 84,000 registros de crédito reales. Los resultados muestran que el modelo de Random Forest fue el más efectivo logrando una precisión del 88% y un AUC de 0.93. Además, utilizaron el método SHAP para identificar las variables que más influían en la predicción, encontrando que la edad y los años de experiencia laboral son factores claves que disminuyen el riesgo de incumplimiento.

Este estudio demuestra la capacidad predictiva de modelos avanzados como Random Forest y XGBoost. Además, también resalta el uso de SHAP, que permite comprender por qué un cliente puede representar mayor o menor riesgo. Esto nos refuerza el uso de estos modelos en contextos como Rena Ware, donde la transparencia en las decisiones de crédito puede fortalecer la gestión del riesgo.

Aplicaciones prácticas en el contexto colombiano

Dentro de las investigaciones del contexto colombiano, SalazarVergara (2022) desarrolló un modelo predictivo orientado al proceso de otorgamiento de microcréditos utilizando algoritmos de clasificación como árboles de decisión, redes bayesianas y k-vecinos más cercanos. El estudio mostró que el algoritmo árboles de decisión alcanzó una precisión del 92,5% siendo el modelo más eficiente entre los evaluados. Las variables más influyentes en la predicción del resultado fueron el historial financiero, los ingresos mensuales, el tipo de empleo y el nivel educativo. Aunque el modelo usado tuvo una precisión importante, este se restringe un poco ya que no contempla algoritmos más avanzados como Random Forest o XGBoost, lo cual podría representar una oportunidad para mejorar la capacidad predictiva en contextos empresariales.

Por otro lado, Salamanca Arias & Benítez Urrea (2018), desarrollaron un modelo estadístico que permitiera estimar la probabilidad de incumplimiento de los clientes en una empresa del sector siderúrgico colombiano. Aplicaron modelos de regresión logística sobre datos internos

de comportamiento de clientes, identificando variables clave como los días de mora, el score de riesgo interno, la concentración de ventas y el número de operaciones activas.

Aunque el estudio se realizó en un contexto industrial, sus resultados y hallazgos son aplicables a empresas comerciales como Rena Ware, que también otorgan créditos directos a sus clientes. Esta investigación demuestra que es posible construir modelos efectivos a partir de datos internos, sin depender exclusivamente del buró de crédito. Asimismo, demuestra que metodologías tradicionales como la regresión logística siguen siendo útiles cuando se busca equilibrio entre precisión y facilidad para interpretar los datos.

Bajo la misma línea de enfoque, se encuentra Angarita Villamizar & Marin Guerra (2019) quienes desarrollaron un modelo predictivo de incumplimiento de pagos en una entidad de microfinanzas colombiana. Aplicaron técnicas de machine learning como Random Forest, XGBoost y regresión logística con penalización Lasso, encontrando que los árboles de decisión presentaron un mejor desempeño predictivo (F1 Score superior al 0.56). Por otro lado, la regresión Lasso obtuvo un mejor equilibrio entre precisión e interpretabilidad, lo cual es fundamental en el sector financiero colombiano dadas las exigencias normativas.

Esta investigación nos muestra cómo al incorporar variables operativas, factores de comportamiento sectorial y técnicas de balanceo de datos, se puede mejorar significativamente la predicción sin sacrificar trazabilidad ni claridad en la toma de decisiones.

Propuestas metodológicas que incorporan modelos híbridos, técnicas explicables o el uso de información alternativa

Desde una perspectiva metodológica alternativa, Pham Thanh & Nguyen Dinh (2022) desarrollaron un modelo híbrido para predecir el cumplimiento de pagos de préstamos, combinando algoritmos genéticos, regresión logística y redes neuronales. Este modelo realiza una selección automática de variables mediante algoritmos evolutivos, lo cual permite reducir la

dimensionalidad sin sacrificar capacidad predictiva. Luego, aplica la regresión logística para establecer una base y finalmente, emplea una red neuronal para refinar la clasificación y mejorar la precisión.

Los resultados mostraron mejoras notables en las métricas relacionadas con precisión y sensibilidad con relación al uso de algoritmos individuales, lo cual destaca la ventaja de integrar enfoques complementarios.

Por otro lado, Martínez Álvarez et al. (2024) analiza modelos de calificación crediticia que incluye información alternativa como redes sociales, psicometría, open banking, y registros de dispositivos móviles. La investigación identificó que estas fuentes, al ser evaluadas bajo criterios técnicos y normativos, pueden mejorar la predicción de los modelos crediticios y ampliar el acceso al financiamiento para sectores como las MiPymes. Aunque el informe no desarrolla modelos estadísticos específicos, ofrece una base metodológica sólida y herramientas para analizar la calidad de los datos. En conjunto, los estudios evaluados, permiten concluir que la predicción del incumplimiento de pagos es un tema ampliamente estudiado en donde se han mostrado mejoras significativas en el rendimiento predictivo, la interpretación de variables relevantes y el tratamiento de bases de datos que también podrían ser aplicables a contextos como Rena Ware.

A partir de la literatura e investigaciones revisadas, se observa un vacío importante en la aplicación de estos modelos al sector comercial no bancario. En este sentido, se destaca la necesidad de incluir estos modelos a compañías como las de Rena Ware de Colombia, quienes otorgan crédito y cuyo apetito de riesgo es mayor al de las entidades financieras tradicionales.

Por lo tanto, y teniendo en cuenta los referentes internacionales y nacionales, podemos concluir que las técnicas principales a estudiar serán la Regresión Logística, por su carácter interpretativo y robusto en contextos locales, y el Random Forest, por su capacidad predictiva y

eficacia en el manejo de desbalance de clases. Estas dos metodologías serán la base técnica para evaluar la incidencia en el incumplimiento de pagos y afectación en la cartera de créditos de Rena Ware en el período 2022-2025.

b. Marco teórico

El crédito es un instrumento financiero clave en la dinámica económica, pero su uso conlleva el riesgo inherente de impago, lo que exige un análisis riguroso de los factores asociados a dichos incumplimientos. (Superintendencia Financiera de Colombia, 2002).

Es así, como el siguiente marco teórico busca reunir las bases normativas, económicas y metodológicas que explican el riesgo de incumplimiento crediticio, articulando las teorías que respaldan la pertinencia de las variables revisadas en este estudio y justifican la idoneidad de la base de datos analizada.

I. Fundamentos del Crédito y Riesgo de Crédito

Definición de crédito y su función económica

El crédito es un mecanismo financiero que permite a las personas, empresas e instituciones acceder a bienes, servicios o recursos monetarios con la promesa de pago en el futuro, usualmente acompañado de una contraprestación en forma de intereses. Esta herramienta crediticia es uno de los mecanismos más importantes del sistema financiero actual ya que facilita el traslado de recursos hacia los inversionistas o consumidores, permitiendo así una asignación más eficiente del capital en la economía.

En términos macroeconómicos, el crédito cumple una función estratégica en tres dimensiones clave: dinamizar el consumo, apalancar la inversión productiva y expandir el acceso a oportunidades económicas, tanto para los hogares como para las empresas. Desde la perspectiva

de los hogares, el crédito permite suavizar el consumo a lo largo del tiempo, enfrentar contingencias o financiar decisiones de alto impacto como la compra de vivienda o educación superior. En el caso de las empresas, facilita el financiamiento de operaciones, innovación, expansión y mejora de competitividad.

Este proceso conlleva al riesgo crediticio, definido como la posibilidad de que un deudor incumpla con sus obligaciones de pago, ocasionando pérdidas para el acreedor (Superintendencia Financiera de Colombia, 2002).

Según la Superintendencia Financiera de Colombia (2002) el riesgo de crédito no solo se refleja en el incumplimiento del pago, sino también en el deterioro progresivo de la capacidad de pago que tiene la persona que adquiere la deuda. Por lo tanto, la correcta gestión de cartera implica identificar, medir, y controlar el riesgo en todas las etapas del crédito: originación, seguimiento, reestructuración y recuperación. Es así como el desarrollo de modelos para evaluar el riesgo ha ganado protagonismo en las entidades financieras, los entes reguladores y analistas de riesgo, especialmente en escenarios de incertidumbre económica o inestabilidad financiera en el país.

Clasificación de carteras según la Superintendencia Financiera de Colombia.

En Colombia, la Superintendencia Financiera de Colombia (2002) clasifica la cartera crediticia con base en el destino del crédito. Dentro de los tipos de carteras de crédito, conforme su clasificación se encuentran cuatro modalidades: Vivienda, dirigida a personas naturales para compra de vivienda propia. Consumo, dirigida a personas naturales para la compra de bienes durables y servicios. Microcrédito, dirigida a unidades productivas de menor escala y Comercial, Orientada al financiamiento empresarial.

Para el caso de la empresa Rena Ware, su categoría de otorgamiento de créditos es crédito de consumo, entendido como lo indica la Superintendencia Financiera de Colombia, aquellos créditos otorgados a personas naturales cuyo objeto sea financiar la adquisición de bienes de consumo o el pago de servicios para fines no comerciales o empresariales, independientemente de su monto (Superintendencia Financiera de Colombia, 2002).

Al igual que existen categorías de tipos de crédito, existen unas categorías de clasificación de riesgo de crédito las cuales, buscan ubicar a los deudores en un grupo específico considerando su perfil y probabilidad de incumplimiento en sus obligaciones crediticias. Dichas categorías, según son:

Tabla 2. *Categorías de riesgo de crédito de consumo según los meses de mora*

Categoría	Nivel de Riesgo	Descripción	Meses de Mora
A	Riesgo normal	Créditos con adecuada estructuración y atención. La información financiera refleja capacidad de pago adecuada del deudor.	0 meses
B	Riesgo aceptable, superior al normal	Créditos atendidos con aceptable protección, pero con debilidades que podrían afectar el pago si no se corrigen oportunamente.	>1 y hasta 2 meses
C	Riesgo apreciable	Créditos con insuficiencia en la capacidad de pago del deudor, que comprometen el cumplimiento de la obligación en los términos pactados.	>2 y hasta 3 meses
D	Riesgo significativo	Créditos con alta probabilidad de incumplimiento. Presentan características agravadas respecto al riesgo apreciable.	>3 y hasta 6 meses

E	Riesgo de incobrabilidad	Créditos considerados incobrables, con deterioro evidente de la capacidad de pago.	Más de 6 meses
----------	--------------------------	--	----------------

Fuente: Adaptado de (Superintendencia Financiera de Colombia, 2002)

En cuanto a las causas que generan más influencia y afectan en el incumplimiento de pagos, a lo largo de la literatura se han identificado variables relevantes que pueden agruparse en las siguientes categorías:

- **Variables sociodemográficas**, como la edad, el género y el nivel educativo, permiten identificar perfiles de riesgo asociados a condiciones personales del solicitante.
- **Variables económicas**, como los ingresos mensuales y la estabilidad laboral, se vinculan con la capacidad de pago y la probabilidad de cumplir o no con las obligaciones.
- **Historial crediticio**, que incluye aspectos como el puntaje, número de créditos que el solicitante ha tenido o la existencia de morosidades anteriores, refleja la experiencia financiera y el comportamiento del cliente.
- **Características del crédito**, como el monto otorgado, el plazo pactado y la tasa de interés, determinan la carga financiera que será asumida por el deudor y su sensibilidad ante factores externos como la inflación o el desempleo.

En el contexto colombiano, las dinámicas que maneja el sector financiero y el constante riesgo al que se enfrentan las compañías que apalancan su operación mediante el otorgamiento de créditos, ha impulsado el uso de la analítica de datos como una herramienta estratégica dentro de su proceso crediticio. Se destaca el uso de modelos de clasificación dada su capacidad de predecir y automatizar procesos de evaluación asociados al incumplimiento de pagos.

II. Enfoques teóricos aplicados al análisis del Riesgo Crediticio

Los enfoques teóricos que se relacionan en el presente trabajo de investigación se incluyen con el fin de ofrecer un marco conceptual que permitan comprender el riesgo de crédito y el incumplimiento de pagos desde la literatura económica y organizacional. Estas teorías no serán validadas empíricamente en el desarrollo del presente estudio, ni constituyen hipótesis a desarrollar. Las presentes teorías, cumplen una función explicativa y contextual que orienta la interpretación de los resultados y la selección de variables que fueron utilizadas en el análisis aplicado.

Teoría de la información asimétrica

La teoría de la información asimétrica, propuesta por Akerlof plantea que en muchas transacciones económicas una de las partes posee más información relevante que la otra, lo cual puede distorsionar los resultados del mercado. En el contexto del crédito, En su obra “The Market for Lemons”, Akerlof demuestra cómo esta asimetría puede llevar a decisiones riesgosas dado que una parte (normalmente el vendedor o el solicitante) oculta o no revela información crítica para el comprador o prestamista (Akerlof, 1970).

Al momento de requerir un crédito, los solicitantes conocen mejor que el prestamista sus intenciones, ingresos reales o disposición para pagar. Este riesgo que se toma al momento de otorgar el crédito puede llevar a que se otorguen préstamos a quienes tienen menor capacidad o voluntad de pago.

Para el caso de Rena Ware, que otorga créditos directos a sus clientes con un apetito de riesgo más alto que una entidad financiera, la información asimétrica sirve como marco de referencia para comprender de manera conceptual, el riesgo asociado a la falta de información

completa sobre el comportamiento de pago de los clientes y por ello resulta fundamental desarrollar modelos de evaluación y clasificación crediticia que reduzcan este riesgo.

Teoría del Agente y el Principal (Problema de Agencia)

La teoría del agente y el principal surge de la economía de contratos y se enfoca en las relaciones donde una parte (el principal) delega una tarea a otra (el agente), esperando que esta actúe en su interés (Jensen et al., 1976). El problema surge cuando los intereses del agente no están completamente alineados con los del principal y hay asimetría de información que impide supervisar eficazmente su comportamiento. Esto genera costos de agencia, como el riesgo de que el agente actúe de forma oportunista. (Jensen et al., 1976).

En el contexto crediticio, el prestamista actúa como el principal, pues es quien otorga el crédito y el prestamista actúa como agente, quien se compromete a cumplir con los pagos futuros. No obstante, una vez el préstamo es desembolsado, el prestamista no puede controlar cómo el cliente usará el dinero y si cumplirá con la promesa de pago. Si el agente actúa en función de su beneficio personal, por ejemplo, incumpliendo intencionalmente o priorizando otros gastos, se produce un problema de agencia.

En la empresa Rena Ware de Colombia, la teoría del agente y el principal permite comprender desde un enfoque conceptual, las dinámicas de riesgo que se encuentran presentes en el otorgamiento de crédito directo a los clientes ya que, Rena Ware al financiar directamente a sus clientes y una vez aprobado el crédito, no tiene ningún mecanismo de control para el seguimiento al comportamiento del cliente más allá de la gestión operativa y gestión de cobranza. En este sentido, los modelos de clasificación desarrollados en esta investigación no tienen como objetivo validar dicha teoría, sino apoyar la gestión del riesgo crediticio mediante el análisis de

información contribuyendo a una toma de decisiones más informada en el proceso de aprobación y seguimiento de los créditos.

Teoría de la Selección Adversa y Riesgo Moral

La selección adversa y el riesgo moral son dos manifestaciones clave del problema de la información asimétrica en los mercados crediticios, ampliamente desarrolladas por autores como Akerlof (1970) y Stiglitz & Weiss (1981). Los dos conceptos muestran cómo la desigualdad en el acceso a la información puede deteriorar la calidad de la cartera.

La selección adversa ocurre en la fase previa a la concesión del crédito ya que los prestamistas no pueden conocer completamente el perfil de riesgo de los solicitantes, aquellos con mayores probabilidades de incumplimiento tienden a ser los que más buscan financiamiento. Si no se aplican mecanismos adecuados de estudio y perfilamiento del cliente, puede conducir a que la empresa apruebe créditos a clientes de algo riesgo, elevando la morosidad del portafolio.

Por su parte, el riesgo moral se presenta una vez otorgado el crédito, cuando el comportamiento de pago del deudor puede modificarse una vez obtiene el préstamo. Al no tener seguimiento sobre el cliente, este podría priorizar otras deudas o dejar de pagar confiando en que el prestamista no podrá sancionarlo de manera inmediata.

Esta teoría al igual que la mencionada anteriormente, son muy relevantes para comprender desde un enfoque conceptual, los riesgos a los que se enfrenta Rena Ware de Colombia ya que el apetito de riesgo al financiar directamente a sus clientes es más alto que el sector financiero y no tiene ningún mecanismo de control más allá de la gestión operativa y gestión de cobranza.

Teoría de la Segmentación de Mercados (Marketing y Análisis de Clientes)

La teoría de la segmentación de mercados, propuesta por Smith (1956) plantea que los consumidores no son homogéneos y que pueden agruparse en segmentos con características, necesidades o comportamientos similares. Para el análisis crediticio, la segmentación permite clasificar a los clientes según variables demográficas, socioeconómicas y de comportamiento financiero, facilitando la gestión para ofrecer productos de acuerdo con sus límites de crédito.

Para nuestro caso de estudio, los clientes de Rena Ware, se pueden agrupar en los siguientes segmentos:

- **Segmentos demográficos:** Agrupación de clientes por edad, género o ubicación geográfica.
- **Segmentos conductuales:** Se pueden identificar patrones según comportamiento de pago (no morosos vs. morosos), frecuencia de compra, canal de adquisición (presencial o en línea).
- **Segmentos crediticios:** Evaluando variables como historial de pagos, nivel de ingresos estimado o recurrencia de mora.

Es así, como la teoría de la segmentación de mercados se utiliza como un marco conceptual que apoya la interpretación de los resultados y la comprensión de patrones de comportamiento para nuestro caso de estudio.

3. Capítulo III. Diseño metodológico.

a. Enfoque y alcance de la investigación

La presente tesis de investigación tiene un enfoque cuantitativo, con alcance explicativo y predictivo ya que utiliza datos numéricos y técnicas estadísticas para probar hipótesis y predecir fenómenos observados. Hernández Sampieri et al. (2014). En este sentido, este estudio se

desarrollará a partir de información estructurada y tabulada lo que permitirá realizar mediciones precisas, establecer relaciones entre variables y evaluar el desempeño de los modelos con métricas objetivas. En primera instancia, se busca caracterizar el comportamiento de la cartera de la empresa Rena Ware con el fin de identificar patrones relevantes en los datos y posterior, se determinará la relación y/o asociación entre un conjunto de variables independientes de carácter demográfico y crediticio y el incumplimiento de pagos de créditos otorgados por Rena Ware a sus clientes en un periodo específico. Lo anterior, determinando la correlación entre variables y haciendo uso de modelos de clasificación.

Así mismo, en el plano epistemológico, el estudio tiene un enfoque positivista el cual, según Dulal (2025) se caracteriza por hacer hincapié en una realidad compartida y en principios universales haciendo uso de métodos cuantitativos y la observación empírica, tal como es el caso de estudio. Además, en los enfoques positivistas se utilizan datos empíricos y la aplicación de métodos estadísticos para validar hipótesis.

De igual forma, el estudio es no experimental, transversal y correlacional. Es no experimental porque no se alteran las variables de forma directa, sino que se presentan tal como se muestran en los registros del dataset. Es transversal porque el análisis se realiza sobre datos correspondientes a un corte transversal específico (créditos originados en el periodo comprendido entre enero 2022 y junio 2025) y cada crédito tiene un horizonte de observación en función de su estado de cumplimiento de pago desde la fecha de originación hasta la fecha de corte de extracción del dataset del CRM de la compañía (noviembre de 2025). Es correlacional porque se busca establecer el grado de influencia y/o asociación entre las variables independientes demográficas y crediticias y la variable dependiente que es el incumplimiento de pagos.

Finalmente, el estudio también incluye la evaluación de modelos de clasificación a través de métricas de precisión, sensibilización, entre otras, lo cual, permitirá determinar cuál tipo de modelo se ajusta en mayor medida a un contexto de no pago de los créditos.

b. Datos de la investigación

La muestra de análisis está conformada por 11.860 créditos otorgados por Rena Ware en el periodo 2022 a junio 2025 y la unidad de análisis corresponde a cada crédito asociado a un cliente. Esto se debe a que un mismo cliente puede tener uno o varios créditos y la información financiera y demográfica registrada no necesariamente es la misma, sino que corresponde a las condiciones que tenía el cliente al momento de la aprobación del crédito (por ejemplo, variaciones en ingresos, nivel de endeudamiento o puntajes de riesgo). En consecuencia, cada crédito constituye un registro único y completo de análisis que consolida tanto las características del cliente como las condiciones propias del crédito. Estos datos, fueron proporcionados por la compañía y provienen del Buró de crédito y proceso de vinculación de cada cliente a Rena Ware. El tratamiento sobre los datos se realizó considerando:

1) Extracción, transformación y cargue de la base de datos:

Al dataset inicial suministrado, extraído en formato Excel desde el sistema CRM de Rena Ware se le aplicaron procesos de anonimización con el fin de evitar la individualización de los clientes, eliminando variables de identificación como nombres, apellidos, direcciones y emails.

Luego, en la etapa de transformación, se normalizaron las variables para garantizar uniformidad en los registros estandarizando los tipos de datos por categorías (valores monetarios, fechas, variables binarias, etc.). Adicionalmente, se llevó a cabo un proceso de depuración que incluyó un análisis exploratorio inicial para identificar y dar tratamiento a los valores faltantes o

nulos en las variables demográficas y crediticias. Como se muestra en la Tabla 3, los valores faltantes no se concentraban en un único tipo de variable (demográfica o crediticia) ni en una única variable en específico, sino que estaban distribuidos en las diferentes variables de intereses. Las variables con mayores niveles de faltantes o nulos fueron el Estrato con un 13,76% y el % cuota inicial con 7,04%. Lo anterior sugiere que, la ausencia de información no responde a una característica particular de un tipo de variable, sino que, obedecen principalmente a limitaciones en la disponibilidad de información del buró de crédito y/o del proceso de vinculación, asociadas a clientes con historial financiero incompleto o no bancarizados, más que a un patrón derivado del comportamiento de incumplimiento. Adicional a ello, la totalidad de registros nulos correspondían a la clase mayoritaria No morosos.

Tabla 3. Distribución de valores nulos del dataset inicial por variable de modelamiento

Variable	% valores nulos o faltantes
Puntaje	2,18%
% cuota inicial	7,04%
Tasa efectiva anual	0%
Saldo a financiar	0,18%
Ingresos totales	3,17%
Edad	1,8%
No. De cuotas	0%
Total créditos abiertos	0,44%
Estrato	13,76%
Genero	2,87%

Fuente: Elaboración propia.

Por lo anterior y considerando que dichas variables son fundamentales para el modelamiento del riesgo de incumplimiento y que los registros nulos correspondían a créditos No morosos, tipo mayormente presente en la muestra, se optó por eliminar los registros incompletos en su totalidad, evitando así introducir información artificial al dataset que afecte la identificación de patrones reales de pago y sin que esto incurra en pérdida de información relevante dado que, la presencia de nulos fue generalizada en diferentes variables, no solo para la variable de interés, por lo que, estos registros no aportarían mayor información relevante para el modelamiento. En consecuencia, como muestra la Tabla 4, no se eliminó ningún registro de la clase minoritaria (Morosos) conservando la totalidad de los 3.629 registros por contar con información completa en las variables seleccionadas para el análisis. La distribución final de los 11.860 registros corresponde a un 30.6% para Morosos y 69.4% para No Morosos.

Tabla 4. Composición de la base de datos antes y después de la depuración.

Estado del crédito	Registros iniciales	Registros finales	% final
Morosos	3.629	3.629	30.6%
No moroso	13.670	8.231	69.4%
Total	17.299	11.860	100%

Fuente: Elaboración propia.

Por otro lado, y a forma de contexto, se relaciona a continuación el diccionario de variables de la base de datos inicial proporcionada por la empresa:

Tabla 5. *Diccionario de datos de base de datos inicial proporcionada por la empresa.*

Nombre de la variable	Descripción	Tipo de dato	Naturaleza
Tax ID	Número de identificación del cliente	Integer	Identificación numérica
Tax Type	Identificador único del cliente. CC / CE	Integer	Identificación numérica
Order ID	Código de referencia único de la orden del crédito	Integer	Identificación numérica
Order Source	Tipo de orden: Finance Order	String	Categórica nominal
Contrato	Número o código del contrato de crédito	String	Categórica nominal
Created Date	Fecha en que se genera la orden de crédito	Datetime	Temporal
Antigüedad	Días desde que se creó el crédito hasta el 08/11/2025.	Integer	Numerica
Ciudad	Ciudad de residencia del cliente	String	Categórica nominal
Departamento	Departamento de residencia	String	Categórica nominal
Status	Estado actual del crédito (activo, castigado, pagado, slow)	String	Categórica nominal
Incumplimiento	Estado cumplimiento del crédito (moroso, no moroso)	String	Categórica nominal
Valor compra	Monto total de la compra financiada	Float	Numérica continua
Cuota inicial	Valor de la cuota inicial pagada	Float	Numérica continua
Saldo a Financiar	Monto pendiente después de la cuota inicial	Float	Numérica continua
% cuota inicial	Porcentaje de cuota inicial sobre el valor total	Float	Numérica continua
No. de cuotas	Número de cuotas pactadas	Integer	Numérica discreta

Tasa Efectiva Anual	Tasa de interés anual aplicada al crédito	Float	Numérica continua
Días de Mora	Número de días en mora acumulados	Integer	Numérica discreta
Rango de mora	Segmentación interna de altura de mora en días (0, 1-30, 31-60, 61-90, 91-120, 121-150, 151-180, >180)	String	Categórica ordinal
Saldo total	Saldo total vigente del crédito	Float	Numérica continua
Puntaje	Score crediticio asignado al cliente	Integer	Numérica discreta
Quanto Valor Mínimo	Valor mínimo de la cuota “Quanto”	Float	Numérica continua
Reconocer Más	Variable crediticia adicional (ej. monto reconocido adicional)	Float	Numérica continua
Año_Nacimiento	Año de nacimiento del cliente	Integer	Temporal (año)
Valor_Cuota	Valor de la cuota periódica pactada	Float	Numérica continua
Edad	Edad del cliente (derivada de Año_Nacimiento)	Integer	Numérica discreta
Total Cred Abiertos	Número de créditos abiertos en el sistema	Integer	Numérica discreta
Ingresos_Totales	Ingresos declarados por el cliente	Float	Numérica continua
Apalancamiento	Relación entre deuda e ingresos	Float	Numérica continua
Acierta	Variable de score externa (ej. Buró)	Integer	Numérica discreta
Quanto	Variable asociada al modelo de riesgo Quanto	Float	Numérica continua
Estrato	Clasificación del inmueble de residencia asociada a	Integer	Categórica ordinal

	condiciones de vivienda y acceso a servicios públicos.		
Genero	Genero del cliente (femenino, masculino)	String	Categórica nominal

Fuente: Elaboración propia a partir del dataset.

Ahora bien, para el dataset final usado en el proceso de modelamiento, se seleccionaron las siguientes variables las cuales fueron identificadas como las de mayor relevancia en relación con el riesgo de incumplimiento (Tabla 6):

Tabla 6. Variables finales utilizadas en el dataset para modelamiento.

Nombre de la variable	Tipo de dato
Saldo a Financiar	Float
% cuota inicial	Float
No. De cuotas	Integer
Tasa Efectiva Anual	Float
Puntaje	Integer
Edad	Integer
Total, Cred Abiertos	Integer
Ingresos Totales	Float
Estrato	Integer
Genero	Integer
Incumplimiento	Integer

Fuente: Elaboración propia a partir del dataset.

2) Análisis exploratorio, descriptivo y correlacional de los datos

El análisis exploratorio y descriptivo se realizó de manera inicial para comprender el comportamiento general de la muestra e identificar patrones relevantes. En primera instancia se aplicaron técnicas de estadística descriptiva, incluyendo medidas de tendencia central como lo es la media, mediana, moda, y de dispersión con el fin de caracterizar las variables demográficas y crediticias. Asimismo, se emplearon visualizaciones gráficas y diagramas de dispersión que permitan identificar patrones generales, distribuciones atípicas y patrones en los datos. De igual manera, se examinaron las correlaciones entre variables cuantitativas de tipo demográfico y crediticio para establecer relaciones con la variable dependiente (incumplimiento de pago). En el análisis también se evaluó la consistencia de los datos para detectar posibles sesgos y verificar el equilibrio de clases entre créditos cumplidos y en mora. Posteriormente, se seleccionaron las variables para aplicar técnicas de balanceo y modelado.

3) Balanceo de la base de datos:

Teniendo en cuenta que el dataset cuenta con información total de la cartera en el periodo específico, lo que incluye créditos Morosos y No morosos, fue necesario balancear los datos con el fin de no sesgar el aprendizaje de los modelos hacia los datos mayoritarios.

Para ello, se usó la ponderación de clases (`class_weight`), usada en algoritmos como la Regresión Logística y el Random Forest. Esta técnica ajusta el peso de cada clase en el proceso del entrenamiento para dar mayor relevancia a los créditos en mora y penalizando en mayor medida los errores de clasificación. De esta forma, se busca equilibrar la influencia de ambas clases en el proceso de aprendizaje y mejorar la capacidad de detección de incumplimientos. (Hosmer et al, 2013).

También se aplicó el sobre muestreo mediante SMOTE (Synthetic Minority Oversampling Technique) una técnica ampliamente documentada en la literatura, SMOTE genera ejemplos sintéticos de la clase minoritaria a partir de combinaciones de vecinos más cercanos, lo que incrementa su representación en el conjunto de entrenamiento sin necesidad de duplicar registros existentes. Este procedimiento permite que el modelo aprenda patrones más representativos de la clase de incumplimiento y, en consecuencia, mejore su capacidad de predicción en escenarios de desbalance. (Batista et al, 2004)

Así mismo, además de las técnicas de ponderación de clases y sobremuestreo se emplearon métodos de submuestreo aleatorio (Random Undersampling) con el fin de reducir el desbalance entre los créditos cumplidos (mayoritarios dentro de la base) y créditos morosos (minoritarios). Esta técnica consiste en eliminar aleatoriamente una parte de los registros de la clase mayoritaria, logrando una proporción más equilibrada entre ambas categorías. Su principal ventaja radica en la simplicidad y en la reducción del tiempo de entrenamiento del modelo, al trabajar con un conjunto de datos más compacto.

Por otra parte, se usó la técnica de sobremuestreo aleatorio (Random Oversampling), una técnica que replica aleatoriamente registros de la clase minoritaria hasta alcanzar un equilibrio proporcional con la clase mayoritaria. A diferencia del submuestreo, este enfoque conserva la totalidad de los datos originales y permite que el modelo aprenda de un conjunto más representativo de la clase de interés (morosos). Es importante mencionar que, al duplicar tanto volumen de datos, puede aumentar el riesgo de sobreajuste si no se aplica de manera adecuada.

La incorporación de estas estrategias de balanceo permitirá mitigar el desbalance de clases y aumentar la robustez de los modelos, garantizando una evaluación más justa de su desempeño en la predicción del incumplimiento de pagos.

c. Técnicas y modelos

Para el análisis descriptivo, exploratorio y correlacional se usaron las siguientes técnicas:

Estadística descriptiva

La estadística descriptiva nos permitió caracterizar la cartera de créditos y las variables demográficas y crediticias de los clientes mediante medidas de tendencia central y medidas de dispersión. Estas herramientas fueron cruciales para entender la distribución de los datos, visualizar patrones generales de existir y detectar valores atípicos.

- **Medidas de tendencia central:** permitiendo identificar valores representativos de las variables, como el ingreso promedio, la edad de los clientes o el valor medio de las cuotas, brindando un panorama general del comportamiento de los clientes.
- **Medidas de dispersión:** incluye indicadores como la desviación estándar, la varianza y el rango, los cuales permiten analizar la variabilidad de los datos y detectar diferencias significativas entre grupos de clientes.
- **Visualizaciones gráficas:** Histogramas, diagramas de caja y gráficos de dispersión apoyarán la interpretación de las características de la base de datos y complementarán el análisis descriptivo con una representación visual de tendencias y dispersiones.

Análisis correlacional

Se aplicarán coeficientes de correlación Pearson o Spearman. Esto dependerá del origen y naturaleza de las variables, de tal manera que se puedan identificar correlaciones entre las variables independientes y su asociación con el incumplimiento de pagos.

- **Estadístico de Pearson:** el coeficiente de correlación de Pearson mide la fuerza y dirección de la relación lineal entre dos variables cuantitativas continuas. Su valor oscila entre -1 y +1, donde los valores cercanos a ± 1 indican una fuerte relación lineal y los valores cercanos a 0 indican ausencia de correlación lineal. Este coeficiente parte del supuesto de que las variables presentan distribución normal, homocedasticidad y una relación estrictamente lineal. (Levin & Fox, 2014)
- **Coeficiente de Correlación de Spearman:** es una medida no paramétrica que se utiliza para evaluar la fuerza y dirección de la relación monotónica entre dos variables. A diferencia de Pearson, Spearman no requiere que los datos cumplan supuestos de normalidad ni linealidad ya que se basa en los rangos de las observaciones en lugar de los valores originales. Este coeficiente resulta útil cuando las variables son de naturaleza ordinal o cuando los datos cuantitativos no cumplen los criterios estadísticos necesarios para aplicar métodos paramétricos. (Siegel & Castellan, 2006)

Una vez identificada la correlación entre variables, se procederá a la selección de aquellas con mayor influencia en el incumplimiento de pagos y/o asociación de variables, para reducir la dimensionalidad y mejorar el rendimiento de los modelos.

Modelos predictivos de clasificación

Los modelos de clasificación son algoritmos de aprendizaje supervisado utilizadas para predecir categorías con base en el cumplimiento de unas características específicas que distinguen a los grupos de categorías entre sí. En el contexto de riesgo crediticio puede ser el cumplimiento o incumplimiento de pagos. Entre los modelos de clasificación más comunes se encuentran:

Regresión Logística

La regresión logística se utiliza para predecir un resultado binario a partir de un conjunto de variables, que pueden ser continuas, discretas o una combinación de ambas y está dada por la formula:

$$\log \left(\frac{P(Y = 1)}{1 - P(Y = 1)} \right) = B_0 + B_1X_1 + B_2X_2 + \dots + B_kX_k$$

Donde,

Y: variable dependiente binaria (ej. incumplimiento = 1, cumplimiento = 0).

P(Y=1): probabilidad de que ocurra el evento (por ejemplo, que un cliente incumpla).

$\frac{P(Y=1)}{1-P(Y=1)}$: razón de probabilidades del evento.

log: logaritmo natural (función logit).

B_0 : intercepto del modelo (constante).

B_1, B_2, B_k : coeficientes asociados a cada variable explicativa.

X_1, X_2, X_k : variables independientes (predictoras).

Fuente: Adaptado de (Hosmer et al., 2013)

Random Forest

Es un modelo que combina varios árboles de decisión contruidos a partir de diferentes subconjuntos aleatorios del conjunto de entrenamiento y de las variables predictoras. Cada árbol hace una predicción independiente y la decisión final se selecciona por mayoría de votos (clasificación) o promedio (regresión). (L. Breiman, 2001) Su formulación puede expresarse como:

$$\hat{y} = \text{mod} h_1(x), h_2(x), \dots, h_t(x)$$

Donde,

$h_t(x)$: es la predicción del árbol t-ésimo entrenado con un subconjunto aleatorio de datos y variables.

x : Representa la observación o registro del cliente, es decir, el conjunto de variables independientes como edad, ingreso, historial, etc.

$\hat{y}$: es la clase final predicha (por ejemplo, "incumple" o "cumple") determinada por mayoría de votos entre los árboles.

t : Es el número total de árboles que componen el modelo.

Fuente: Adaptado de (L. Breiman, 2001)

Con la aplicación de los modelos anteriores se espera determinar las variables incidentes en el incumplimiento de pagos y predecir el incumplimiento de un crédito con base en las características demográficas y crediticias del cliente.

De igual forma para el procesamiento, análisis y modelado de los datos se emplearán herramientas como Google Colab y/o Python, utilizando librerías como pandas, scikit-learn, matplotlib, seaborn y statsmodels.

Técnica de optimización: Método de la rejilla

El método de rejilla es una técnica de optimización que evalúa todas las combinaciones posibles de hiperparámetros dentro de un espacio finito definido. Su principal fortaleza es la

simplicidad y la garantía de explorar completamente dicho espacio, lo que permite identificar la mejor configuración dentro de los valores establecidos (Bergstra & Bengio, 2012)

Este método se utiliza frecuentemente para ajustar modelos Random Forest porque permite explorar de forma sistemática los hiperparámetros que controlan la complejidad y el comportamiento del conjunto de árboles. En este algoritmo, aspectos como el número de árboles, la profundidad máxima o el número mínimo de muestras para dividir un nodo influyen directamente en la capacidad predictiva y en el riesgo de sobreajuste. La búsqueda en rejilla evalúa cada combinación posible de estos parámetros mediante validación cruzada, lo que garantiza una exploración completa del espacio de búsqueda definido (L. Breiman, 2001).

Técnicas de evaluación de desempeño de modelos de clasificación

Los modelos serán evaluados utilizando métricas como, exactitud (Accuracy), precisión (precision), sensibilidad (Recall), F-1 score y AUC-ROC (área bajo la curva ROC), con el fin de seleccionar el modelo más eficaz.

- **Exactitud (Accuracy):** fracción de predicciones que el modelo clasifica correctamente, es decir, el número de aciertos dividido entre el número total de predicciones (James et al., 2013)
- **Precisión (Precision):** mide la proporción de instancias correctamente identificadas como positivas entre todas las que el clasificador predijo como positivas. (Kuhn & Johnson, 2013)

- **Sensibilidad (Recall):** la sensibilidad, también conocida como tasa de verdaderos positivos o Recall, corresponde a la proporción de instancias positivas que fueron correctamente identificadas por el clasificador. (Han et al., 2012)
- **Especificidad (Specificity):** es la proporción de observaciones negativas que son correctamente identificadas como negativas por el modelo, complementando la medida de sensibilidad (Shmueli et al., 2020).
- **F-1 score:** métrica de evaluación que combina la precisión y la sensibilidad en un solo valor mediante la media armónica de ambas. Se utiliza principalmente en problemas de clasificación en especial cuando las clases están desbalanceadas y es necesario equilibrar el costo de falsos positivos y falsos negativos. (Provost & Fawcett, 2013)

La fórmula matemática es:

$$F1 = 2 * \frac{(Precisión * Sensibilidad)}{(Precisión + Sensibilidad)}$$

Un **F1 = 1** indica el mejor desempeño posible (precisión y sensibilidad perfectas).

Un **F1 = 0** indica el peor desempeño, donde el modelo falla en ambas métricas.

Valores intermedios del F1 reflejan el balance entre la capacidad del modelo para identificar correctamente las instancias positivas (sensibilidad) y para no clasificar erróneamente instancias negativas como positivas (precisión) (Han et al., 2012).

Área bajo la curva ROC (AUC-ROC): el área bajo la curva ROC (AUC) resume el rendimiento de un clasificador a través de todos los posibles umbrales de decisión; un AUC cercano a 1 indica un fuerte poder discriminativo, mientras que un valor de 0.5 indica azar (Hastie et al., 2009).

Matriz de confusión: Matriz que representa tabularmente los falsos positivos y falsos negativos, verdaderos positivos y verdaderos negativos, lo que permite calcular las métricas anteriores de desempeño y analizar de manera integral el rendimiento del modelo.

Ahora bien, considerando el contexto de negocio de riesgo de crédito, no todas las métricas de desempeño de los modelos de clasificación tienen el mismo impacto. Para el caso de la empresa Rena Ware, el riesgo al que están más expuestos es aprobar créditos a clientes que finalmente incumplan sus obligaciones, lo que generará pérdidas financieras y costos de cobranza. Lo anterior traducido a las métricas de desempeño quiere decir que, es necesario priorizar métricas como el Recall de la clase morosa dado que es la métrica que indica la capacidad del modelo para identificar correctamente a los clientes morosos. Un Recall bajo implicaría dejar pasar clientes de alto riesgo, lo cual, en ocasiones resulta más costoso que rechazar de forma preventiva a clientes que podrían llegar a pagar correctamente.

No obstante, el modelo de negocio de empresas de ventas a crédito en gran medida es rentable por el mismo hecho de otorgar los créditos y asumir riesgos, es por ello por lo que, debe haber equilibrio entre mantener el riesgo minimizado y al mismo tiempo aceptar a algunos clientes, aunque tengan probabilidades de incumplimiento. Por lo anterior, el F1-score cobra relevancia en estos contextos de negocio como métrica de referencia porque permite evaluar el equilibrio entre la sensibilidad (Recall) y la Precisión, evitando modelos muy restrictivos que afecten la aprobación de créditos rentables. Mantener este enfoque al analizar y evaluar los modelos de clasificación permitirá alinear el desempeño del modelo con el apetito de riesgo de la empresa.

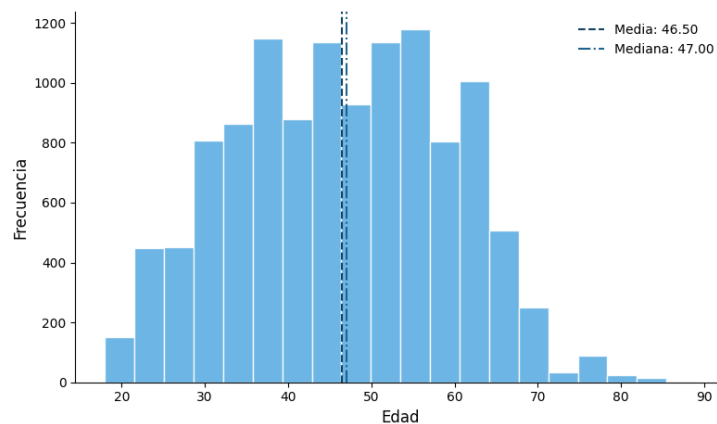
Si bien ambos modelos de clasificación Regresión logística y Random Forest son comúnmente usados en el contexto financiero, es fundamental evaluarlos considerando las métricas anteriores para determinar el modelo que ofrece el mejor equilibrio y desempeño entre dichas métricas.

4. Capítulo IV. Resultados

a. Caracterización de la cartera de créditos

La muestra de análisis conformada por 11.860 créditos otorgados a clientes de Rena Ware tienen las siguientes características:

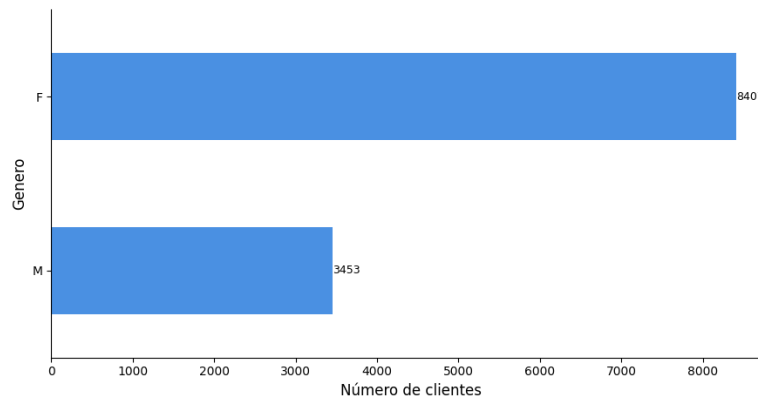
Ilustración 1. Distribución de los clientes por edad



Fuente: Elaboración propia a partir del dataset.

La Ilustración 1 muestra el comportamiento de la variable edad en el dataset de clientes, donde se evidencia una mayor concentración en los rangos de edad entre los 38 y 67 años por lo cual, el promedio de edad de clientes es de 46 años aproximadamente. Estos resultados evidencian que los clientes son principalmente adultos en etapas productivas de su ciclo de vida y con capacidad adquisitiva y estabilidad laboral, lo que facilita su acceso a otorgamientos de créditos.

Ilustración 2. Distribución de los clientes por género



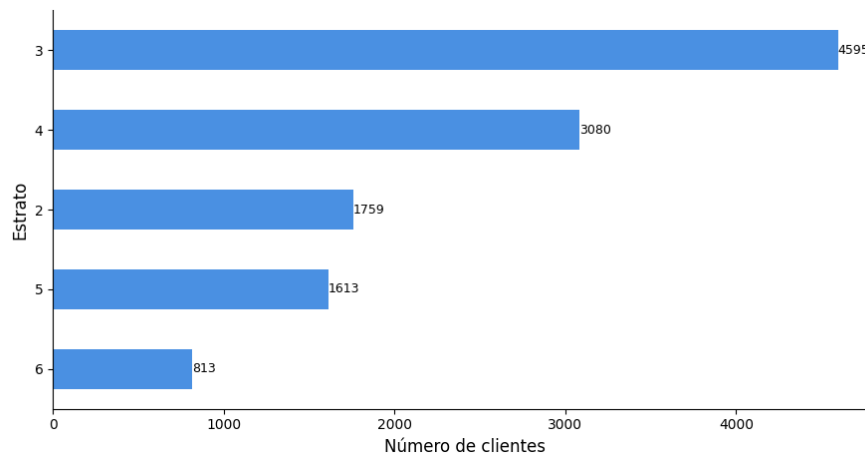
Fuente: Elaboración propia a partir del dataset.

En cuanto a la distribución de clientes por género se encuentra que, como indica la Ilustración 2 la mayoría de los clientes son mujeres con 8.407 créditos, es decir el 71%, mientras que los hombres representan 3.453 registros, es decir el 29%. Esta distribución es coherente con la dinámica del sector de venta directa en Colombia y con los tipos de clientes de Rena Ware en el sentido que las mujeres son la mayor parte de la fuerza comercial (Acovedi, 2022). Además, el tipo de producto que comercializa Rena Ware (utensilios de cocina premium) se encuentra en mayor medida relacionado con las necesidades de compra de mujeres.

Por otro lado, respecto al estrato, variable que nos permite aproximar condiciones de vivienda y acceso a servicios de los clientes y que puede estar asociada indirectamente con la capacidad adquisitiva de los mismos, en la Ilustración 3 se observa que la mayoría de los clientes pertenecen a los estratos 3 y 4 con 4.595 y 3.080 registros respectivamente, lo que en conjunto representa cerca del 65% de la cartera total. Esta concentración muestra que Rena Ware focaliza su nicho de clientes en los estratos medios de la población colombiana quienes cuentan con la capacidad adquisitiva suficiente para acceder a productos premium, pero al mismo tiempo recurren al crédito como medio para financiar la adquisición de estos. Los estratos 5 y 6 con

1.613 y 813 registros respectivamente, tienen una menor participación, lo cual sugiere que en estos segmentos existe menor necesidad del crédito otorgado por la empresa para acceder a los productos que ofrece.

Ilustración 3. Distribución de los clientes por estrato.

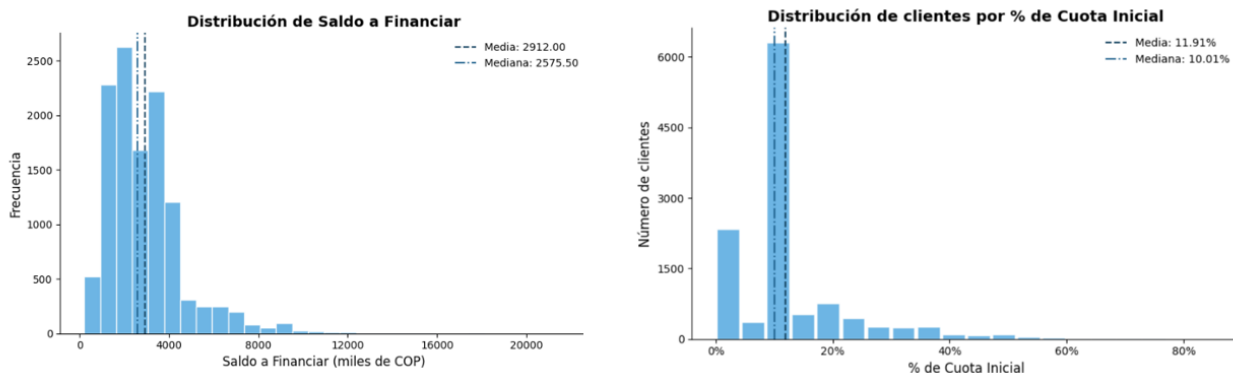


Fuente: Elaboración propia a partir del dataset.

En cuanto a la distribución del saldo a financiar en los créditos otorgados y el % de la cuota inicial de dichos créditos, la Ilustración 4 muestra que los montos de saldo a financiar están concentrados en valores inferiores a 4 millones de pesos, con una media de \$2,912,000 COP.

Por otro lado, el porcentaje de cuota inicial de la compra oscila hasta máximo el 20%, siendo la media el 11.91%. Lo anterior refleja que, el valor total de la compra en promedio sería 3.2 millones de pesos, lo cual, se considera montos relativamente moderados y, que el saldo a financiar y el % de cuota inicial son variables que podrían presentar correlación y a su vez, ser relevantes frente al incumplimiento considerando que, los créditos de mayor monto a financiar y con cuotas iniciales bajas, implican un mayor compromiso financiero para el cliente y aumenta el riesgo de posible incumplimiento de la obligación.

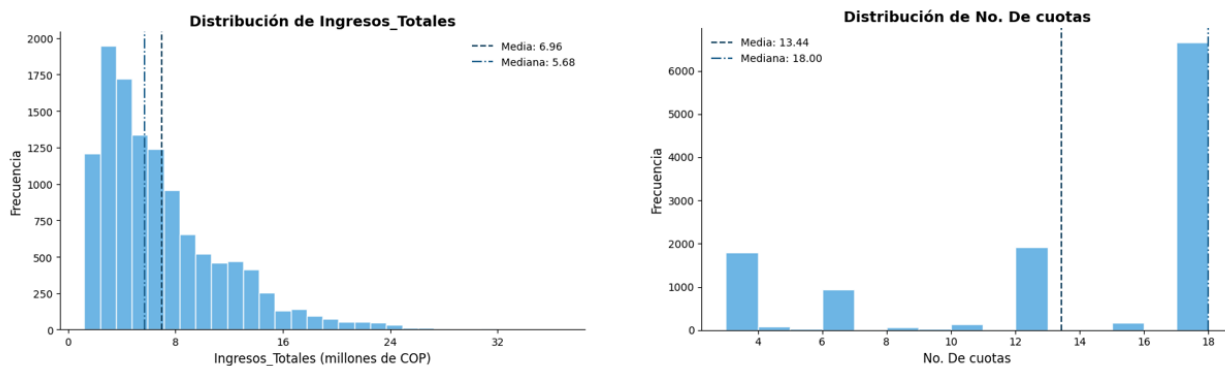
Ilustración 4. Distribución de los clientes por saldo a financiar y % de cuota inicial



Fuente: Elaboración propia a partir del dataset.

En la misma línea de revisión de variables crediticias, en la Ilustración 5 se observa la distribución de los clientes según sus ingresos totales y el número de cuotas a las que se difiere el crédito.

Ilustración 5. Distribución de clientes por ingresos totales y número de cuotas del crédito



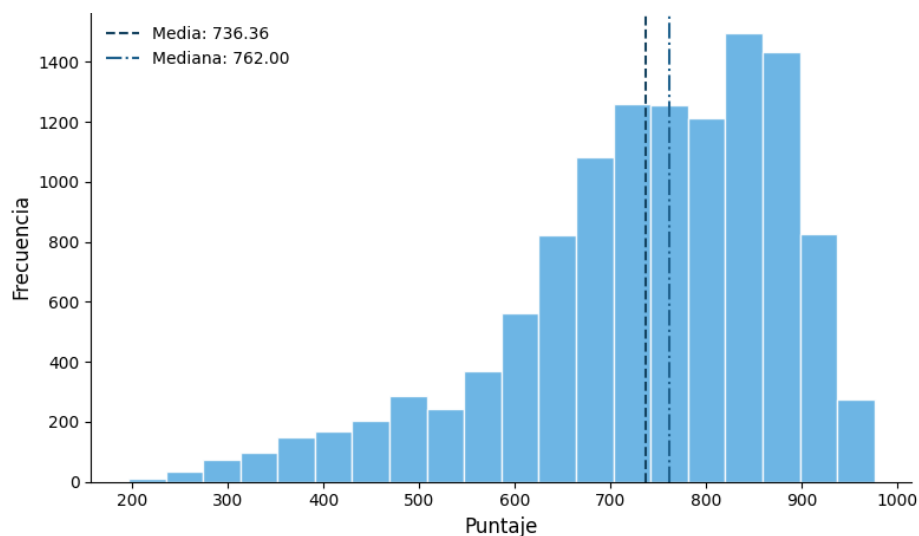
Fuente: Elaboración propia a partir del dataset.

Como se observa, los ingresos totales declarados por los clientes al buró de crédito tienen una concentración entre los 3 y 8 millones de pesos mensuales con una media de aproximadamente 7 millones por cliente. Así mismo, en cuanto al número de cuotas pactadas en los créditos se observa que normalmente se otorgan a plazos de 12 y 18 cuotas.

En términos de riesgo crediticio, los ingresos como el número de cuotas son variables relevantes en el análisis del incumplimiento teniendo en cuenta que, los ingresos del cliente se relacionan con su capacidad de pago y el plazo del crédito influye en el flujo de egresos mensuales del cliente, así como aumenta el riesgo de incumplimiento.

Otra variable relevante para analizar es el comportamiento del puntaje de score crediticio que tienen los clientes asignado en el buró de crédito. Este indicador que va desde 150 a 950 puntos explica que tan bueno es el cliente con el cumplimiento de sus obligaciones financieras y cuál es su capacidad de endeudamiento.

Ilustración 6. Distribución de clientes por puntaje crediticio del buro

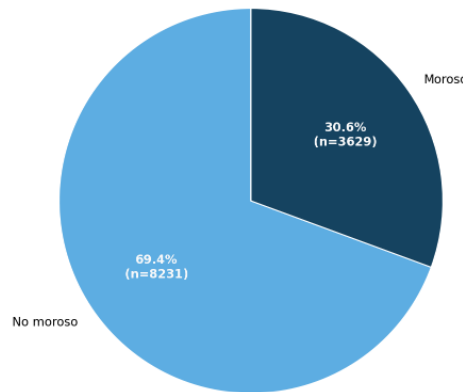


Fuente: Elaboración propia a partir del dataset.

En la anterior Ilustración se observa una concentración de los clientes en puntajes de score superior a los 600 puntos y una media de 736 puntos. Esto sugiere que la mayoría de los clientes cuenta con un perfil de riesgo favorable. No obstante, la presencia de clientes con puntajes bajos demuestra que la compañía mantiene un apetito de riesgo moderado al incluir dentro de su portafolio de clientes perfiles más riesgosos a impago. En este sentido, el puntaje se

convierte en una de las variables más relevantes a analizar como determinante ante el incumplimiento puesto que, este indicador crediticio permite diferenciar entre clientes con alta y baja probabilidad de incumplimiento.

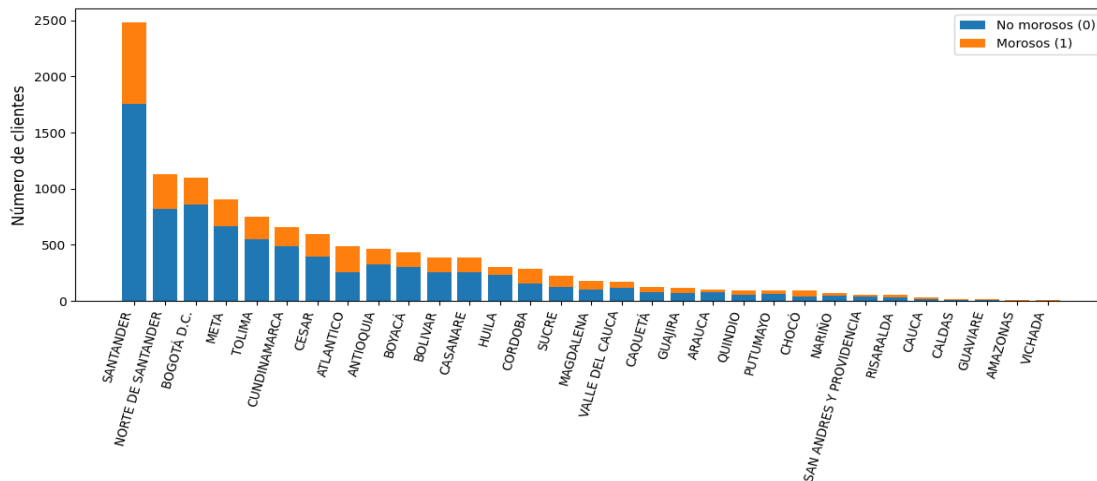
Ilustración 7. Distribución de clientes por incumplimiento



Fuente: Elaboración propia a partir del dataset.

En cuanto a la distribución de clientes por incumplimiento, la Ilustración 7 muestra que el 69.4% de créditos corresponden a No morosos mientras que el 30.6% son Morosos. Este resultado refleja que, si bien la gran mayoría de créditos se mantienen al día, el segmento moroso, aunque es minoritario representa un riesgo financiero para Rena Ware. De igual forma, la proporción es consistente con las cifras reportadas oficialmente por la empresa en las que, a diciembre de 2024, el 87,51% de los créditos se encontraban vigentes o con mora menor a 30 días. No obstante, es importante mencionar que la caracterización fue realizada con la base de datos sin manipular en relación con el balanceo de clases.

Ilustración 8. Distribución de clientes por departamento de ubicación e incumplimiento.



Fuente: Elaboración propia a partir del dataset.

Ahora bien, en cuanto a la distribución por departamento de residencia, en la Ilustración 8 se evidencia una alta concentración en el departamento de Santander con cerca de los 2.500 registros (21,07% de la cartera total) seguido por Norte de Santander con 1.126 (9.49%) y Bogotá D.C. con 1.099 (9.2%). Lo anterior indica que la cartera crediticia de la compañía se concentra principalmente en regiones centrales y con una trayectoria consolidada en la operación comercial, destacándose Santander como el departamento predominante. Así mismo, la menor participación se presenta en departamentos con niveles más bajos de desarrollo económico y alejados de ciudades principales. De igual forma, se destaca que el departamento del Atlántico es el que tiene una distribución casi igualitaria para Morosos y No morosos, siendo del 47.96% para Morosos (236 créditos) y 52.04% No morosos (256 créditos).

b. Relación entre variables independientes

Frente a la relación existente entre las variables independientes demográficas y crediticias de tipo numérico se encuentra que, como se evidencia en la Ilustración 9 existe una correlación fuerte positiva entre las variables Estrato e Ingresos totales del 0.76 lo cual infiere que ambas

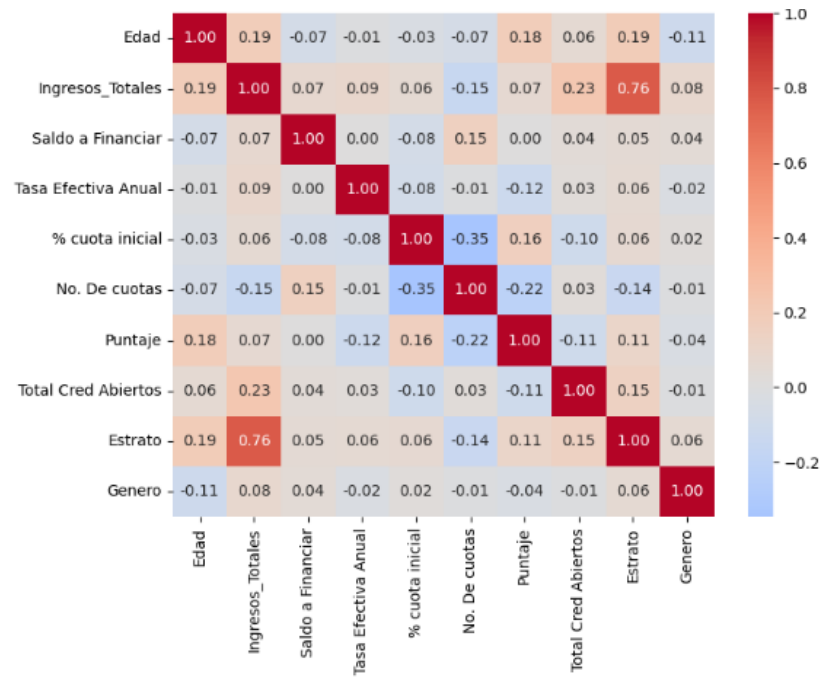
variables pueden estar asociados a condiciones socioeconómicas generales del entorno residencial y económico de los clientes. De igual forma, existe correlación moderada positiva entre el Estrato y Edad (0.19), Estrato y Total créditos abiertos (0.15) y, Estrato y Puntaje (0.11). Lo anterior evidencia la congruencia de relación entre variables socioeconómicas y de condiciones de entorno de vivienda.

Así mismo, hay correlación moderada positiva entre Ingresos totales y Edad de 0.19 y de 0.23 con Total de créditos abiertos. Esto indica que, a mayor edad, mayor ingreso y a mayor ingreso, mayores créditos son otorgados a los clientes.

En cuanto al Saldo a financiar y % de Cuota inicial, existe una correlación negativa baja de - 0.08 y entre el % de cuota inicial y No. De cuotas hay una relación moderada negativa de - 0.35. Lo cual tiene sentido en relación con que, a mayor porcentaje de cuota inicial, es menor el saldo a financiar y a mayor % de cuota inicial, es requerido menor número de cuotas para la financiación.

Ahora, en cuanto a las correlaciones más relevantes de la variable Puntaje se tiene que, son moderadas negativas con No. De Cuotas de -0.22, Total créditos abiertos de -0.11 y Tasa efectiva anual de -0.12, lo que explica que, el puntaje es determinante en las condiciones de financiación del crédito. Así mismo, Puntaje tiene correlación moderada positiva con % de Cuota inicial de 0.16, con Edad de 0.18 y con Estrato de 0.11. Lo anterior justifica que a mayor edad y estrato mejor el puntaje crediticio del cliente, es decir que, normalmente personas con mayor estrato son de edad mayor y, por tanto, tiene el poder adquisitivo para que le sean otorgados más créditos.

Ilustración 9. Mapa de correlación entre variables numéricas.



Fuente: Elaboración propia a partir del dataset.

En cuanto al resto de las variables analizadas no se observan relaciones significativas entre ellas. No obstante, a continuación, se resume el ranking de variables independientes que afectan en mayor medida a la variable objetivo incumplimiento:

Ilustración 10. Ranking de variables más importantes sobre el incumplimiento

Variable	Importancia
Puntaje	0.177853
% cuota inicial	0.145409
Tasa Efectiva Anual	0.141815
Saldo a Financiar	0.121980
Ingresos_Totales	0.118725
Edad	0.103534
No. De cuotas	0.074207
Total Cred Abiertos	0.064328
Estrato	0.035180
Genero	0.016969

Fuente: Elaboración propia a partir del dataset.

De la anterior Ilustración podemos concluir que más de la mitad de influencia en el incumplimiento está determinada por variables de tipo crediticio (Puntaje, % cuota inicial, Tasa Efectiva anual, Saldo a Financiar e Ingresos totales) siendo las variables demográficas Edad, Estrato y Genero de menor relevancia.

c. Implementación y evaluación de desempeño de los modelos de clasificación

Regresión Logística y Random Forest

Para la predicción del incumplimiento de pagos de créditos de clientes de la empresa Rena Ware se implementaron y evaluaron 2 modelos de clasificación: Regresión Logística y Random Forest, cada uno entrenado sin balanceo y posterior bajo diferentes técnicas de balanceo: Class_weight (ponderación de clases), RUS (Random Undersampling), ROS (Random Oversampling) y Smote (Synthetic Minority Oversampling Technique). Los resultados obtenidos fueron:

Ilustración 11. Métricas generales de los modelos de clasificación con técnicas de balanceo: sin balanceo, class_weight, rus, ros, Smote

=== TABLA B – UMBRAL FIJO = 0.5 (ordenada por F1 → ROC → Recall → Precision → Accuracy) ===

Modelo	Tecnica	Umbral	Accuracy	Precision	Recall	F1	ROC_AUC	TN	FP	FN	TP
RandomForest	rus	0.5	0.747470	0.567338	0.736915	0.641102	0.833195	1238	408	191	535
RandomForest	ros	0.5	0.794688	0.699499	0.577135	0.632453	0.840460	1466	180	307	419
RandomForest	smote	0.5	0.787521	0.683775	0.568871	0.621053	0.829876	1455	191	313	413
RandomForest	none	0.5	0.802698	0.780435	0.494490	0.605396	0.842086	1545	101	367	359
RandomForest	class_weight	0.5	0.802698	0.793182	0.480716	0.598628	0.844658	1555	91	377	349
LogisticRegression	smote	0.5	0.679174	0.484105	0.734160	0.583470	0.760949	1078	568	193	533
LogisticRegression	class_weight	0.5	0.679595	0.484461	0.730028	0.582418	0.761613	1082	564	196	530
LogisticRegression	ros	0.5	0.679174	0.483930	0.725895	0.580716	0.761536	1084	562	199	527
LogisticRegression	rus	0.5	0.680860	0.485608	0.720386	0.580144	0.760941	1092	554	203	523
LogisticRegression	none	0.5	0.739882	0.643045	0.337466	0.442638	0.761129	1510	136	481	245

Fuente: Elaboración propia a partir del dataset.

Los resultados evidencian que los modelos alcanzaron una exactitud (Accuracy) entre 0.73 y 0.80 siendo el puntaje más alto para el modelo Random Forest class_weight y none (sin técnica de balanceo) con un valor de 0.8026. Sin embargo, realizando un análisis integral de las

métricas y en el contexto de riesgo de crédito, un valor alto de Accuracy no es suficiente para determinar que un modelo tiene mejor desempeño frente a otro. Para ello, debemos analizar y priorizar otras métricas como lo son el F-1 quien nos da el equilibrio del modelo entre Precision (de los predichos como morosos, cuántos realmente lo son) y Recall (proporción de morosos correctamente identificados, lo que minimiza riesgo de pérdida real), así como también el ROC_AUC quien mide la capacidad general del modelo de discriminar entre Recall o tasa de verdaderos positivos y tasa de falsos positivos.

Considerando lo anterior, los modelos que presentan mejores métricas de desempeño a nivel global son:

Ilustración 12. Métricas de los 3 modelos de clasificación con mejor desempeño global: Random Forest rus, ros y Smote

Modelo	Tecnica	Accuracy	Precision	Recall	F1	ROC_AUC
RandomForest	rus	0.747470	0.567338	0.736915	0.641102	0.833195
RandomForest	ros	0.794688	0.699499	0.577135	0.632453	0.840460
RandomForest	smote	0.787521	0.683775	0.568871	0.621053	0.829876

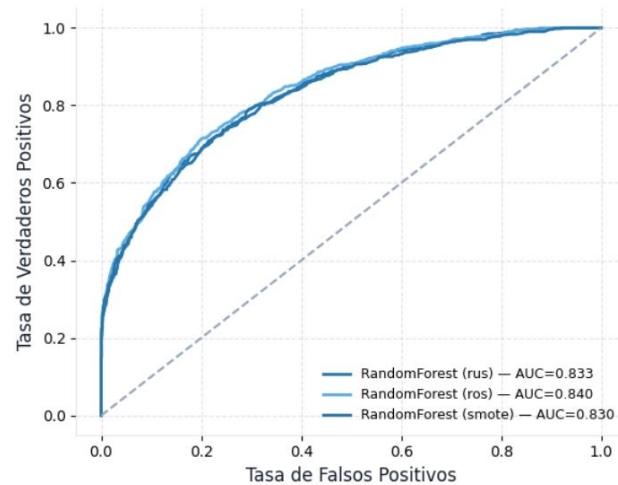
Fuente: Elaboración propia a partir del dataset.

Como se observa en la Ilustración 12, el modelo de clasificación Random Forest fue el que obtuvo mejores resultados de desempeño frente a la Regresión Logística para el contexto de riesgo de crédito. De igual forma, las técnicas RUS (Random Undersampling) y ROS (Random Oversampling) predominaron en las métricas de desempeño.

Finalmente, en cuanto a las curvas ROC, como se evidencia en la Ilustración 13 confirma que los tres modelos mejores alcanzan un rendimiento moderado-alto (AUC entre 0.82 y 0.84) siendo el rus y ros los modelos con valor más alto de AUC (0.83 y 0.84 respectivamente). En conclusión, para modelos de riesgo de crédito, el mejor equilibrio entre sensibilidad y precisión lo ofrece el Random Forest con balanceo rus, ros y Smote. De igual forma, los modelos sin balanceo o con class_weight son susceptibles a tener sesgo hacia los No Morosos por sus Recall

bajos (entre 0.48 y 0.49) lo que quiere decir que su capacidad para detectar Morosos reales es inferior al 50%.

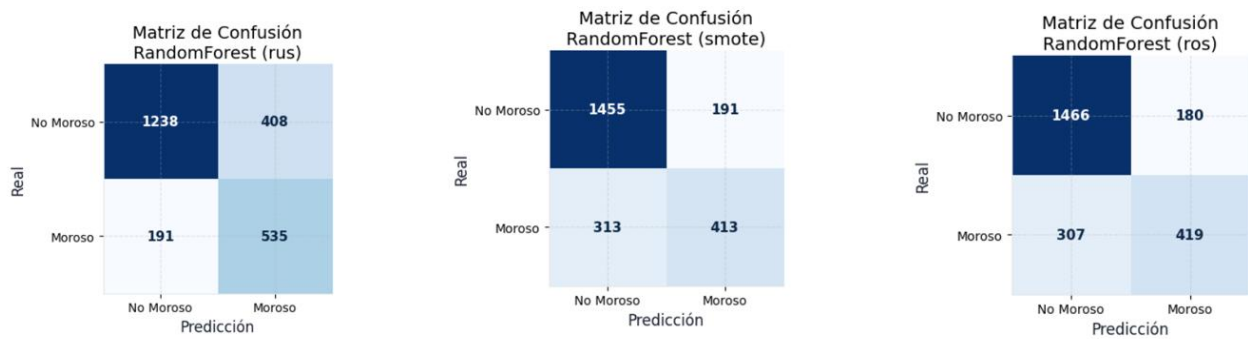
Ilustración 13. Curvas ROC de los 3 modelos con mejor desempeño: Random Forest rus, ros y Smote



Fuente: Elaboración propia a partir del dataset.

Ahora, en cuanto a las matrices de confusión de los tres mejores modelos se tiene:

Ilustración 14. Matrices de confusión de los 3 modelos con mejor desempeño: Random Forest rus, ros y Smote



Fuente: Elaboración propia a partir del dataset.

Las matrices de confusión permiten visualizar la calidad de la predicción de los modelos, dando como resultados:

- El modelo **Random Forest RUS** es el mejor modelo clasificando los clientes Morosos (535) por lo que es agresivo en su selección, no obstante, esto lo logra sacrificando la Precision por lo que deja ir a clientes buenos por clasificarlos como Morosos (408 falsos positivos).
- El modelo **Random Forest Smote** es más equilibrado entre Recall y Precision puesto que, aunque clasifica menos Morosos que el rus (413), tiene menos falsos positivos (191).
- Finalmente, el **Random Forest ROS** podría decirse que es el más conservador porque tiene una mayor precisión (0.69) y al mismo tiempo equilibrio entre Morosos (419) y No Morosos (1466) clasificados correctamente.

En conclusión, cualquiera de los tres modelos evaluados (Random Forest RUS, ROS y Smote) serian adecuados para la clasificación de riesgo de crédito, sin embargo, la elección del modelo óptimo dependerá del apetito de riesgo de cada compañía. En el caso de Rena Ware por su apetito de riesgo alto, se recomienda el modelo Random Forest Smote por ser menos radical y más abierto al riesgo que el RUS y ROS. Esto permite flexibilidad a la hora de la aprobación de nuevos créditos sin exponer en gran medida la cartera de la compañía.

Modelos Random Forest (RUS, ROS y Smote) con método de la rejilla

Con el fin de validar si hay mejoras en las métricas de desempeño de los tres modelos Random Forest determinados como los mejores anteriormente, se procedió a aplicar el método de la rejilla para optimizar los hiperparámetros del algoritmo y encontrar mejores resultados para las métricas. Los resultados fueron los siguientes:

Tabla 7. Comparación de desempeño de los modelos Random Forest antes y después de aplicado el método de la rejilla

Modelo	Optimización	Accuracy	Precision	Recall	F1	AUC
RF RUS	Sin rejilla	0.7474	0.5673	0.7369	0.6411	0.8331
RF RUS	Con rejilla	0.7559	0.5781	0.7493	0.6527	0.8345
RF ROS	Sin rejilla	0.7946	0.6994	0.5771	0.6324	0.8404
RF ROS	Con rejilla	0.7572	0.5845	0.7149	0.6431	0.8335
RF Smote	Sin rejilla	0.7875	0.6837	0.5688	0.6210	0.8298
RF Smote	Con rejilla	0.7799	0.6656	0.5647	0.6110	0.8262

Fuente: Elaboración propia.

Como se evidencia en la tabla anterior, la aplicación del método de la rejilla no generó mejoras significativas en el desempeño predictivo de los modelos. En la mayoría de los casos, las métricas AUC y F1 score se mantuvieron similares o presentaron disminuciones marginales. El único modelo que mejoró las métricas a nivel general con la rejilla fue el Random Forest RUS, sin embargo, esta mejora es marginal de aproximadamente un punto en cada métrica, por lo que se puede decir que fue una mejora homogénea y equilibrada, sin generar grandes beneficios respecto a las métricas sin optimización que ya se encontraban en un punto cercano al óptimo. Por su parte, el modelo Random Forest ROS presentó una mejora significativa en la métrica Recall pasando de 0.57 a 0.71 no obstante, desmejoró en proporción similar la Precision pasando de 0.69 a 0.58. Lo anterior indica que, el modelo tuvo un cambio a ser más sensible en la detección de morosos, pero sacrificó la precisión y exactitud aumentando los falsos positivos. Esto en un contexto de riesgo de crédito es positivo siempre y cuando la empresa sea adversa o

con perfil de riesgo bajo a la morosidad, que en el caso de Rena Ware, su apetito de riesgo es más elevado. Finalmente, con el modelo Random Forest Smote, la optimización no representa beneficios y genera una degradación general del modelo.

A nivel general se puede concluir que, los modelos iniciales se encontraban cercanos a un punto óptimo de configuración y que la optimización adicional no aportó beneficios relevantes en la detección del incumplimiento manteniendo equilibrio en precisión y exactitud. Sumando a ello que, para la aplicación de la optimización se aumentó la complejidad del modelo y el consumo computacional, se considera más adecuado el uso de los modelos sin optimización, por su equilibrio entre simplicidad, flexibilidad para realizar ajustes y desempeño.

A continuación, se muestran los resultados de la optimización por cada modelo más a detalle destacando la configuración de los hiperparámetros:

Random Forest RUS optimizado

Ilustración 15. Métricas e hiperparámetros del modelo Random Forest RUS con optimización

clf__max_depth: 20	Accuracy: 0.7559
clf__max_features: sqrt	Precision: 0.5781
clf__min_samples_leaf: 1	Recall: 0.7493
clf__min_samples_split: 10	F1 Score: 0.6527
clf__n_estimators: 600	AUC ROC: 0.8345
	TN=1249, FP=397, FN=182, TP=544

Fuente: Elaboración propia a partir del dataset.

Para el caso de este modelo, como se evidencia en la Ilustración 15 hay una mejora menor en las métricas Accuracy, Precision, Recall y F1 de un punto aproximante para cada una, respecto al modelo sin optimización. En cuanto a los hiperparámetros, indican que el modelo óptimo utiliza un bosque grande de 600 árboles, con una profundidad moderada alta de 20, nodos

que se dividen cuando contienen suficientes observaciones (mínimo 10) y hojas muy específicas (1 muestra). Además, emplea la selección aleatoria de variables típica del Random forest sqrt.

Si bien hubo una mejora esta no es significativa en comparación a las métricas del modelo inicial y el consumo computacional de este método es considerablemente más alto en comparación al primer modelo.

Random Forest ROS optimizado

Ilustración 16. Métricas e hiperparámetros del modelo Random Forest ROS con optimización.

clf__max_depth: 10	Accuracy: 0.7572
clf__max_features: sqrt	Precision: 0.5845
clf__min_samples_leaf: 1	Recall: 0.7149
clf__min_samples_split: 5	F1 Score: 0.6431
clf__n_estimators: 200	AUC ROC: 0.8335
	TN=1277, FP=369, FN=207, TP=519

Fuente: Elaboración propia a partir del dataset.

En el caso del modelo Random Forest ROS optimizado los resultados desmejoraron en comparación al modelo inicial en las métricas Accuracy, Precision y AUC. Únicamente mejoró en las métricas Recall pasando de 0,57 a 0,71 y el F1 de 0,63 a 0,64 (ver Ilustración 16). Este modelo mostró una configuración con árboles de profundidad moderada (10 niveles), selección aleatoria de variables sqrt y parámetros que permiten capturar patrones relevantes sin sobre ajustar, así como un bosque total de 200 árboles. Con esta optimización el modelo se volvió más sensible a la detección de morosos, pero sacrificó sensibilidad y precisión.

Random Forest Smote optimizado

Para el modelo Random Forest Smote optimizado, los resultados muestran:

Ilustración 17. Métricas e hiperparámetros del modelo Random Forest Smote con optimización.

clf_max_depth: 20	Accuracy: 0.7799
clf_max_features: sqrt	Precision: 0.6656
clf_min_samples_leaf: 1	Recall: 0.5647
clf_min_samples_split: 5	F1 Score: 0.6110
clf_n_estimators: 150	AUC ROC: 0.8262
	TN=1440, FP=206, FN=316, TP=410

Fuente: Elaboración propia a partir del dataset.

Se evidencia en la Ilustración 17 que todas las métricas de desempeño Accuracy, Precision, Recall, F1 y AUC en este modelo desmejoraron en comparación del modelo inicial.

A nivel general la comparación entre los modelos Random Forest sin optimización (Ilustración 12) y los modelos en sus respectivas ajustadas mediante método de la rejilla (Ilustraciones 15, 16 y 17) evidencian que, los modelos iniciales ofrecen un desempeño solido considerando las diferentes métricas de desempeño que son competitivas y con buenos resultados destacándose valores de F1 entre 0.62 y 0.64 y AUC superiores a 0.83, lo que refleja su capacidad de clasificar y su equilibrio entre precisión y sensibilidad. Con la optimización, si bien algunos modelos lograron ajustes en parámetros internos, estos cambios no reflejaron mejoras significativas en las métricas de desempeño, al contrario, en la mayoría de los modelos el rendimiento disminuyó, lo que sugiere que los modelos originales ya se encontraban cercanos a un punto óptimo de funcionamiento e hiperparámetros por lo que, se concluye que la optimización adicional no aportó beneficios adicionales relevantes para la predicción del riesgo de incumplimiento.

5. Capítulo V. Discusiones

Dados los resultados obtenidos con los modelos implementados (Regresión Logística y Random Forest) y diferentes técnicas de balanceo (sin balanceo, class_weight, RUS, ROS y

Smote) en donde se evidencia que lograron un desempeño positivo con métricas AUC entre 0.76 y 0.84 y valores de F1-score entre 0.44 y 0.64 se puede observar que, los hallazgos se alinean con la literatura revisada. Estudios previos coinciden en que tanto los modelos tradicionales como los basados en aprendizaje automático pueden ofrecer resultados sólidos, dependiendo del tipo de datos y del balance aplicado.

Autores como (Baesens et al., 2003) realizaron una amplia comparación de algoritmos de clasificación aplicados al score crediticio y concluyeron que, si bien los modelos avanzados como las redes neuronales superan en desempeño a los modelos tradicionales, la Regresión Logística mantiene su utilidad por su claridad en la interpretación de la información. Los resultados obtenidos para el caso de Rena Ware se alinean con este planteamiento ya que, aunque la Regresión Logística no fue mejor modelo que el Random Forest, si es aceptable para el contexto de riesgo de crédito y obtuvo valores AUC de desempeño aproximados de 0.76.

Por su parte (Coşer et al., 2019) compararon diferentes algoritmos aplicados a una base de datos desbalanceada y encontraron que el modelo Random Forest ofreció el mejor rendimiento. En el caso de los modelos implementados en la presente tesis investigativa, el Random Forest none (sin balanceo) obtuvo excelentes resultados, casi cercados a los modelos Random Forest RUS, ROS y Smote, con valores de Accuracy 0.80, Precision 0.78, Recall 0.49, F-1 0.60 y AUC de 0.84. Su diferencia con los tres modelos mejores fue su Recall bajo que no favorece la detección de Morosos. Lo anterior cobra sentido y alineación con el hecho que el desbalance del dataset es de más del 60% hacia registros de No Morosos.

De igual forma (Li & Wu, 2024) evaluaron el rendimiento de modelos de aprendizaje automático encontrando que el Random Forest alcanzó una precisión del 88% y un AUC de 0.93, además de destacar la importancia de variables como edad y experiencia laboral. Los resultados obtenidos en Rena Ware concuerdan con estas conclusiones ya que, el Random Forest fue el

mejor modelo con métricas AUC superiores a 0.82 y Precision promedio de 70%. Así mismo, las variables Edad, Ingresos y Puntaje demostraron una alta-moderada incidencia sobre la probabilidad de incumplimiento. Esto refuerza la consistencia del patrón de variables influyentes en el contexto de riesgo de crédito independiente de la base de datos.

En el contexto colombiano, Echeverri Valdes (2006) destacó la relevancia de variables operativas como los días de mora y valor del crédito para anticipar el incumplimiento, mientras que (Angarita Villamizar & Marin Guerra, 2019) encontraron que, aunque los modelos de árboles como Random Forest superaban en desempeño a los métodos tradicionales, la Regresión Logística ofrecía un equilibrio óptimo entre interpretabilidad y precisión. En la investigación de Rena Ware se observa una tendencia similar dado que, el Random Forest resultó ser superior en detección de Morosos pero la Regresión Logística, aunque tuvo menor precisión, ofrece rendimiento competitivo entre Recall y F-1 lo que puede indicar que es un modelo menos sensible al sobreajuste.

Ahora bien, desde una perspectiva conceptual, esta investigación se puede contextualizar a partir de los enfoques teóricos que fueron revisados dentro del marco teórico los cuales abordan de manera general, las dinámicas que presenta el mercado relacionadas con el riesgo crediticio y el comportamiento de pago de los clientes. Estos enfoques han sido utilizados en la literatura para describir las problemáticas que presentan las compañías y en especial Rena Ware de Colombia, asociadas a la disponibilidad de información, la toma de decisiones bajo incertidumbre de cumplimiento por parte de los clientes y la gestión del riesgo en operaciones de crédito.

Para el caso de Rena Ware de Colombia, el uso de variables demográficas y crediticias, así como la aplicación de modelos de clasificación, permiten identificar de una manera más oportuna los perfiles con mayor probabilidad de incumplimiento.

No obstante, es importante precisar que en el presente estudio los enfoques teóricos revisados solamente cumplen una función de referencia conceptual para contextualizar el fenómeno analizado, sin que los resultados obtenidos pretendan contrastar o validar dichos marcos teóricos. En consecuencia, los hallazgos se interpretan desde un enfoque aplicado, orientado al fortalecimiento de la gestión preventiva del riesgo crediticio y a la toma de decisiones.

Asimismo, la investigación realizada valida también la eficacia del uso de técnicas de balanceo y depuración de variables en bases de datos, considerando que, la eliminación de variables categóricas que no aportaban mucho al desempeño de los modelos (ej. Departamento y Ciudad) y la eliminación de variables por colinealidad (ej. Valor de la compra, Valor de Cuota inicial) permitió mejorar las métricas generales demostrando la importancia de la selección de variables y la reducción de dimensionalidad.

Finalmente, desde el punto de vista del negocio, los resultados reafirman que el modelo Random Forest, específicamente Random Forest con técnica de balanceo Smote, es la alternativa más adecuada para el contexto de Rena Ware de Colombia. Esto, dada su capacidad de detección de morosos sin incurrir en sobreajustes. Su implementación en los procesos de aprobación de contratos permitirá clasificar las solicitudes de crédito de manera preventiva con el fin de identificar los casos con mayor probabilidad de incumplimiento y así, fortalecer la gestión del riesgo y de cartera de la compañía.

Adicionalmente, la aplicación del método de rejilla para optimización de hiperparámetros no represento una mejora significativa en los resultados de los modelos, por el contrario, en varios casos desmejoró las métricas de desempeño y la capacidad computacional requerida para su implementación fue considerablemente elevada en comparación a los modelos sin optimización.

Ahora bien, es importante mencionar que el desempeño de los modelos se basa en la información demográfica y crediticia con la que se disponía a un corte específico, por lo que los resultados pueden verse afectados por variables no observadas y que no fueron incluidas en el análisis, como condiciones macroeconómicas o el comportamiento de pago de los clientes en un rango de tiempo más amplio. En este sentido, futuras investigaciones podrían incorporar este tipo de información como apoyo a la toma de decisiones, con el fin de mejorar y fortalecer la capacidad predictiva de los modelos en escenarios que son cambiantes.

Adicionalmente, los hallazgos obtenidos tras la elaboración de esta investigación pueden resultar útiles para otras compañías del sector real de venta directa que, como Rena Ware, otorgan crédito directo a sus clientes y enfrentan desafíos similares en la gestión de su cartera y el riesgo crediticio. En particular, empresas que cuentan con información limitada de sus clientes, que no se encuentren bancarizados o compañías con procesos de evaluación crediticia poco estructurados podrían beneficiarse de la implementación de modelos de clasificación basados en variables demográficas y crediticias observables.

En este contexto, la experiencia del caso Rena Ware sugiere que el uso de modelos como Random Forest, acompañado de técnicas de balanceo como ROS, RUS y Smote y una adecuada selección de variables, puede apoyar a empresas del mismo sector o similares para la identificación temprana de posibles incumplimientos de clientes y toma de decisiones informadas, con el fin de contribuir a mejorar la identificación de perfiles de riesgo y la gestión preventiva de la cartera.

6. Capítulo VI. Conclusiones

El desarrollo de esta investigación permitió analizar de manera integral las variables que inciden en el incumplimiento de pagos de los créditos otorgados por Rena Ware de Colombia, confirmando que el uso de modelos de clasificación es una herramienta eficaz para prevenir y anticipar riesgos crediticios y fortalecer la gestión de cartera.

A partir de los resultados obtenidos se evidencia que los objetivos planteados en este proyecto se cumplieron satisfactoriamente. La caracterización del comportamiento de la cartera de créditos de Rena Ware se realizó mediante un análisis descriptivo y exploratorio que permitió identificar las principales tendencias demográficas y financieras de los clientes. Se logró identificar que la mayoría de los créditos corresponden a mujeres en rangos de edad entre 38 y 67 años, pertenecientes a estratos medios (3 y 4), con montos financiados inferiores a cuatro millones de pesos y ubicados principalmente en el departamento de Santander, Norte de Santander y Bogotá. Lo anterior es acorde al tipo de producto que Rena Ware comercializa (utensilios de cocina) y permitió comprender la estructura real de la cartera (69.4% No morosos y 30.6% Morosos) y su concentración por perfil de cliente.

Con relación a la implementación de modelos de clasificación para predecir el incumplimiento de pagos y su comparación en desempeño, se aplicaron los modelos de Regresión Logística y Random Forest sobre la base de datos sin balancear y con diferentes técnicas de balanceo (class_weight, RUS, ROS y Smote) dando como resultado que ambos modelos y técnicas lograron métricas aceptables con AUC entre 0.76 y 0.84 y F1-score superiores a 0.44, lo que demuestra su capacidad predictiva y su coherencia con la literatura revisada.

El modelo Random Forest fue sin duda el mejor modelo y las técnicas de balanceo RUS, ROS y Smote son las que tienen mejor equilibrio entre métricas: AUC promedio de 0.835, Recall de 0.628, Accuracy de 0.777 y F-1 score promedio de 0.632. Aunque finalmente para el contexto de Rena Ware, el modelo que mejor se adapta es el Random Forest con balanceo Smote por su capacidad de detección de Morosos sin ser radical y adverso al riesgo.

Por otro lado, con relación a las variables más determinantes del incumplimiento se destacan el Puntaje crediticio en primer lugar, seguido del % de Cuota inicial, Tasa efectiva anual, Saldo a financiar, Ingresos totales, Edad, Numero de Cuotas, Total créditos abiertos, Estrato y Genero. Esto nos indica que además de haber coherencia con la literatura estudiada, las características sociodemográficas y crediticias influyen en el comportamiento de pago de los clientes siendo notorio que las variables demográficas son de menor relevancia que las crediticias. Aunque el estrato tiene cierta relación con el comportamiento de pago, su influencia en comparación a variables estrictamente crediticias es menor y debe interpretarse como una variable contextual asociada a condiciones de vivienda, más no como un indicador directo de ingresos del cliente.

En el aspecto metodológico, la investigación nos permitió validar técnicas efectivas para el tratamiento de bases de datos desbalanceadas, demostrando que, en este contexto, las técnicas Random Oversampling (ROS), Random Undersampling (RUS) y Smote, ofrece mejores resultados que la ponderación por clases (`class_weight`). También, la depuración de variables categóricas no relevantes y variables con colinealidad permitió mejorar las métricas globales de los modelos, confirmando que la selección de variables es clave para fortalecer la precisión y evitar el sobreajuste.

Ahora, en cuanto a las teorías, los hallazgos de la investigación son consistentes con la literatura revisada, al mostrar que los modelos aplicados permiten reducir la incertidumbre que se asocia al comportamiento de los clientes y apoyan la identificación de perfiles de alto riesgo desde un enfoque aplicado, en línea con lo revisado en la literatura sobre información asimétrica, selección adversa y el riesgo moral. De igual forma, estos resultados permiten comprender desde una perspectiva conceptual las dinámicas de riesgo presentes en la relación entre la empresa y sus clientes, en la medida en que se puede apoyar a una gestión preventiva del riesgo y ayudan a la empresa a anticipar acciones de los clientes y reducir los riesgos mediante una gestión preventiva del riesgo crediticio.

Podemos concluir que la presente tesis de investigación logró cumplir los objetivos planteados y demostró que la combinación de técnicas estadísticas, modelos de clasificación y técnicas de balanceo pueden ser una herramienta efectiva para predecir el incumplimiento de pagos y optimizar la gestión de cartera en compañías que otorgan crédito directo y buscan fortalecer sus procesos de evaluación y gestión de cartera., como es el caso de Rena Ware. La metodología propuesta, puede ser adaptada a contextos empresariales con características similares, siempre que se disponga de información básica demográfica y crediticia de los clientes. Los resultados obtenidos confirman que la analítica de datos puede convertirse en un pilar estratégico para empresas que otorgan crédito directo, como Rena Ware de Colombia, al permitir decisiones más informadas, preventivas y oportunas en un entorno empresarial que requiere data para toma de decisiones.

Finalmente, se concluye que la aplicación del método de la rejilla para la optimización de hiperparámetros no siempre garantiza mejoras en el desempeño de los modelos, por lo que su aplicación debe evaluarse en función del costo computacional que implica y del beneficio real obtenido. Lo anterior considerando que en varios casos los modelos Random Forest al aplicar

método de la rejilla, desmejoraron las métricas de desempeño y en el caso de mejora, que fue el modelo Random Forest RUS, la mejora no fue significativa dado que fue de un punto aproximadamente en cada métrica. Lo anterior corrobora que, los modelos Random iniciales están configurados con parámetros apropiados que permite tengan una capacidad discriminativa buena y un equilibrio entre precisión y sensibilidad.

Referencias

- Acovedi. (2022). *Una empresa con Certificación en el Código de Ética*. Acovedi.
<https://acovedi.org.co/companies/rena-ware/>
- Akerlof, G. A. (1970). The Market for “Lemons”: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), 488–500.
<https://doi.org/10.2307/1879431>
- Angarita Villamizar, J. C., & Marin Guerra, R. A. (2019). Modelo de predicción de incumplimiento de pago en el sector de microfinanzas. *Universidad de los Andes. Maestría en Inteligencia Analítica para la Toma de Decisiones*.
- Baesens, B., Van Gestel, T., Viaene, S., Stepania, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, 54, 627–635.
<https://doi.org/10.1057/palgrave.jors.2601545>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.
- Chaabita, R., Ghoujdam, Elkhalfi, O., Zehraoui, K., Elalaoui, H., & Idamia, S. (2024). Consumer credit risk analysis through artificial intelligence: a comparative study between the classical approach of logistic regression and advanced machine learning

techniques. *Cogent Economics and Finance*, 12(1).

<https://doi.org/10.1080/23322039.2024.2414926>

Coşer, A., Mihaela, M. M.-M., & Crişan Albu. (2019). Predictive Models for Loan Default Risk Assessment. *Economic Computation and Economic Cybernetics Studies and Research*, 53(2), 149–165. <https://doi.org/10.24818/18423264/53.2.19.09>

Crouhy, M., Galai, D., & Mark, R. (2000). *A Comparative Analysis of Current Credit Risk Models. Journal of Banking & Finance*.
https://www.academia.edu/29132514/CREDIT_RISK_Credit_risk_modeling_theory_and_applications

Dulal, T. D. (2025). Application of positivism and post positivism approach in contemporary research. *International Journal of Applied Research and Sustainable Sciences (IJARSS)*.

Echeverri Valdes, F. (2006). *Evaluación de modelos para la medición de riesgo de incumplimiento en créditos para una entidad financiera del Eje Cafetero*.

Forbes. (2023, April 11). *Sector de venta directa creció 8% en 2022 en Colombia: generó 6 billones de pesos en transacciones*. Acovedi. <https://acovedi.org.co/sector-de-venta-directa-crecio-8-en-2022-en-colombia-genero-6-billones-de-pesos-en-transacciones/>

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

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

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6th ed.). McGraw-Hill.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.
- Jensen, M. C., Meckling, W. H., Benston, G., Canes, M., Henderson, D., Leffler, K., Long, J., Smith, C., Thompson, R., Watts, R., & Zimmerman, J. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. In *Journal of Financial Economics* (Issue 4). Harvard University Press.
<http://hupress.harvard.edu/catalog/JENTHF.html>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- L. Breiman. (2001). Random Forests. *Machine Learning*, 45, 5–32.
- Levin, J., & Fox, J. A. (2014). *Estadística para las ciencias sociales* (9th ed.). Pearson Educación.
- Li, H., & Wu, W. (2024). Loan default predictability with explainable machine learning. *Finance Research Letters*, 60. <https://doi.org/10.1016/j.frl.2023.104867>
- Martínez Álvarez, J., Restrepo, L., Dussan, A., Rocha, R., & Edelshein, J. (2024). *Modelos de Calificación Crediticia con Información Alternativa: Recomendaciones para su implementación en Colombia*. www.iadb.org

- Pham Thanh, B., & Nguyen Dinh, T. (2022). *Predicción del pago de préstamos mediante un híbrido de algoritmos genéticos, regresión logística y redes neuronales artificiales*.
<https://link.springer.com/bookseries/7899>
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
- Rena Ware International INC. (2022). *Compañía - Conozca Rena Ware*. Rena Ware International INC. <https://colombia.renaware.com/9b45e6a5-e096-4747-a584-b01e3a7d3756/company>
- Salamanca Arias, A., & Benítez Urrea, J. A. (2018). *Estimación de la probabilidad de incumplimiento de créditos para una empresa del sector siderúrgico en Colombia*.
- Salazar Vergara, J. G. (2022). *Diseño de un modelo predictivo para otorgar créditos*. Universidad de Medellín.
- Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for business analytics: Concepts, techniques, and applications in Python*. Wiley.
- Siegel, S., & Castellan, N. J. (2006). *Estadística no paramétrica aplicada a las ciencias de la conducta* (2nd ed.). McGraw-Hill.
- Smith, Wendell R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, 21(1), 3–8.
<https://doi.org/10.1177/002224295602100102>
- Stiglitz, J. E., & Weiss, A. (1981). *Credit Rationing in Markets with Imperfect Information*. 393–410.

Superintendencia Financiera de Colombia. (2002). *Circular Externa 011 de 2002*.

Superintendencia Financiera de Colombia. (2024). *Actualidad del Sistema Financiero Colombiano*.

Wu, Y., & Pan, Y. (2021). Application Analysis of Credit Scoring of Financial Institutions Based on Machine Learning Model. *Complexity Journal*, 2021.

<https://doi.org/10.1155/2021/9222617>