UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Comparación entre los modelos de regresión y árboles de decisión en la estimación de la calificación crediticia en los microcréditos

Trabajo de titulación previa la obtención del
Título en Magister en Estadística Aplicada

Autor: Ángel Delfín Guaraca Daquilema

Tutor: MSc. Pablo Javier Flores Muñoz

Tulcán, febrero 2025

CERTIFICADO DEL TUTOR

Certifico que el estudiante Guaraca Daquilema Ángel Delfín con el número de cédula 0605287572 ha elaborado el trabajo de titulación: “Comparación entre los modelos de regresión y árboles de decisión en la estimación de la calificación crediticia en los microcréditos”.

Este trabajo se sujeta a las normas y metodología dispuestas en la Codificación del Reglamento de Régimen Académico y de Estudiantes de la Universidad Politécnica Estatal del Carchi con RESOLUCIÓN No. 171-CSUP- 2023, por lo tanto, autorizo su presentación para la sustentación respectiva

PABLO JAVIER FLORES MUNOZ

Firmado digitalmente por PABLO JAVIER FLORES MUNOZ
Fecha: 2025.02.12 15:21:08 +01'00'

…….....

MSc. Pablo Javier Flores Muñoz
TUTOR

Tulcán, febrero 2025

AUTORÍA DE TRABAJO

El presente trabajo de titulación constituye un requisito previo para la obtención del título de Magíster en Estadística Aplicada

Yo, Ángel Delfín Guaraca Daquilema, ciudadano ecuatoriano con cédula de identidad número 0605287572 declaro: que la investigación es absolutamente original, auténtica, personal y los resultados y conclusiones a los que he llegado son de mi absoluta responsabilidad.

Firmado electrónicamente por:
ANGEL DELFIN
GUARACA DAQUILEMA

…….....
Ángel Delfín Guaraca Daquilema
AUTOR

Tulcán, febrero 2025

ACTA DE CESIÓN DE DERECHOS DEL TRABAJO DE TITULACIÓN

Yo, Ángel Delfín Guaraca Daquilema declaro ser autor/a de los criterios emitidos en el trabajo de titulación: “Comparación entre los modelos de regresión y árboles de decisión en la estimación de la calificación crediticia en los microcréditos” y eximo expresamente a la Universidad Politécnica Estatal del Carchi y a sus representantes legales de posibles reclamos o acciones legales.

.....

Ángel Delfín Guaraca Daquilema
AUTOR

Tulcán, febrero 2025

AGRADECIMIENTO

Quiero expresar mis profundos agradecimientos a Dios por bendecirme a cumplir este sueño tan anhelado, por dar fuerza y valor en todo el momento. A mis padres Juan Guaraca y María Daquilema por su amor incondicional y su apoyo constante. A mis hermanos, porque han sido mis pilares fundamentales, gracias por enseñarme que los sueños se construye con esfuerzo y perseverancia. Así también, a María Toapanta, por su paciencia y amor. Gracias por estar siempre a mi lado. Su apoyo ha sido sin duda el viento bajo mis alas. A mis amigos, en especial Roxana Mariño por ser refugio en los momentos de estrés y mi motivación en los momentos de desánimo, su amistad ha sido un regalo valioso de Dios que atesoraré por siempre.

A mi tutor, MSc. Pablo Flores quien con su paciencia, conocimiento y experiencia ha guiado en todo el trayecto, Su guía ha sido tan valioso para alcanzar este objetivo muy importante, mis mejores deseos de éxito en su vida profesional, laboral y personal. Que Dios te bendiga Siempre.

Este esfuerzo no solo es mío; es un reflejo de todos ustedes. ¡Gracias por ser parte de este gran sueño!

DEDICATORIA

Dedico este proyecto a Dios, mi guía esencial en todo mi proceso de formación. A mis padres y hermanos, quienes me han ofrecido un apoyo incondicional en cada fase de mi vida, inculcándome valores, sabiduría y éxito, y ayudándome a convertirme en la persona que soy. Quiero también dedicar este trabajo a mis sobrinas, en especial a Cinthya, Samara y Judith, por ser siempre mi inspiración y maravilloso que es vivir con curiosidad y asombro día a día.

ÍNDICE

RESUMEN xii

ABSTRACT xiii

CAPÍTULO I 1

PROBLEMA 1

1.1. Planteamiento del problema 1

1.2. Preguntas de investigación 3

1.3. Objetivos de investigación 3

1.3.1. Objetivo General 3

1.3.2. Objetivos Específicos 3

1.4. Justificación 4

CAPÍTULO II 6

FUNDAMENTACIÓN TEÓRICA 6

2.1. Antecedentes de investigación 6

2.2. Marco teórico 13

2.2.1. Factores influyentes en la medición de la calificación crediticia en microcréditos 13

2.2.2. Estimación de riesgo crediticio 14

2.2.3. Análisis de comparación entre el modelo de regresión y árboles de decisión 27

2.1. Marco Legal 33

CAPÍTULO III 35

METODOLOGÍA 35

3.1. Descripción del área de estudio 35

3.2. Enfoque y tipo de investigación 36

3.3. Definición y operacionalización de variables 37

3.4. Procedimientos 39

CAPÍTULO IV 43
4. RESULTADOS Y DISCUSIÓN 43
4.1. Determinar las variables que influyen en la estimación de la calificación crediticia de los microcréditos de COAC Fernando Daquilema Ltda. en el 2023. 43
4.2. Estimar la calificación crediticia de los microcréditos de la COAC Fernando Daquilema Ltda. a través de la aplicación de los modelos de regresión y de árboles de decisión. 52
4.2.1. Regresión Logística 53
4.2.2. Árboles de decisión 65
4.3. Analizar la calificación crediticia de los microcréditos otorgados en la COAC Fernando Daquilema Ltda. por medio de la comparación de la precisión predictiva y facilidad de implementación entre los modelos de regresión y árboles de decisión. . 73
CONCLUSIONES Y RECOMENDACIONES 78
5.1. Conclusiones 78
5.2. Recomendaciones 80
REFERENCIAS 81
ANEXOS 88

ÍNDICE DE TABLAS

Tabla 1 Matriz de confusión 28
Tabla 2 Definición de variables 37
Tabla 3 Operacionalización de variable independiente 38
Tabla 4 Operacionalización de la variable Dependiente 39
Tabla 5 Recodificación de las variables 40
Tabla 6 Variable Tasa de Interés 45
Tabla 7 Distribución de frecuencia tipo de producto 46
Tabla 8 Distribución de frecuencias tipo de producto discretizada 47
Tabla 9 Distribución de frecuencia modalidad de pago 48
Tabla 10 Distribución de frecuencia modalidad de pago discretizada 48
Tabla 11 Distribución de frecuencia nivel de riesgo 49
Tabla 12 Distribución de frecuencia nivel de riesgo discretizada 51
Tabla 13 Partición de la base de datos 52
Tabla 14 Variables significativas 53
Tabla 15 Significancia de las variables predictoras 55
Tabla 16 Validación del modelo 56
Tabla 17 Parámetros estimados de las variables significativas 59
Tabla 18 Matriz de confusión 62
Tabla 19 Distribución relativa de la variable dependiente 66
Tabla 20. Valor de probabilidad 66
Tabla 21 Frecuencia relativa de variable dependiente 68
Tabla 22 Matriz confusión Arboles de decisión 70
Tabla 23 Pérdida Logarítmica 73
Tabla 24 Métricas de Evaluación 74
Tabla 25 Variables Significativas 76

ÍNDICE DE FIGURAS

Figura 1 Función Logística 19
Figura 2 Estructura de árbol de decisión 26
Figura 3 Ejemplo de Curva ROC 31
Figura 4 Ejemplo de Punto de corte Óptimo 32
Figura 5 Área de estudio 36
Figura 6 Distribución de riesgo de crédito según la calificación 50
Figura 7 Distribución de la variable dependiente 51
Figura 8 Curva ROC, datos de prueba 64
Figura 9 Punto de corte óptimo 65
Figura 10 Curva ROC con diferentes valores de probabilidad 67
Figura 11 Gráfica de modelo árbol de decisión 68
Figura 12 Curva ROC modelo árbol de decisión 72
Figura 13 Comparación de modelos 75

ÍNDICE DE ECUACIONES

Ecuación 1 Algoritmo de Regresión Logística 17
Ecuación 2 Ecuación Lineal 17
Ecuación 3 Función lineal de Regresión Logística 17
Ecuación 4 Función Logit 17
Ecuación 5 Logit o predictor lineal 17
Ecuación 6 Predictor lineal, multiplicado por exponencial 18
Ecuación 7 Regresión Logística 18
Ecuación 8 Razón entre probabilidades 20
Ecuación 9 Razón entre Odds 20
Ecuación 10 Estadístico de Wald 21
Ecuación 11 Estadístico Wald para n grande 21
Ecuación 12 Estadístico Chi- cuadrado 22
Ecuación 13 Chi-cuadrado de Pearson 22
Ecuación 14 Estadística Desvianza 23
Ecuación 15 Desvianza para Regresión Logística 23
Ecuación 16 Exactitud del modelo 29
Ecuación 17 Tasa de Error de clasificación 29
Ecuación 18 Tasa de verdaderos positivos 29
Ecuación 19 Tasa de verdaderos negativos 29
Ecuación 20 Tasa de falsos positivos 30
Ecuación 21 Tasa de falsos negativos 30
Ecuación 22 Pérdida Logarítmica 30
Ecuación 23 Área bajo la curva -AUC 31
Ecuación 24 Índice de Youden 32
Ecuación 25 Estadístico de prueba para pruebas individuales 54
Ecuación 26 Estadístico G 56
Ecuación 27 Modelo de Regresión Logística 57

ÍNDICE DE ANEXOS

Anexo A Certificado del Abstract por parte de Centro de Idioma 88
Anexo B Modelo de árbol de decisión 89
Anexo C Resumen de modelo de árbol de decisión..... 89

RESUMEN

La presente investigación evalúa la calificación crediticia de los microcréditos otorgados por la Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda. en el 2023, por medio de la comparación de criterios de desempeño entre el modelo de regresión y arboles de decisión a fin de seleccionar el modelo con mejor precisión predictiva. La investigación tiene un enfoque cuantitativo, de tipo correlacional y explicativo. Los modelos se compararon empleando métricas de precisión, la matriz de confusión, curva ROC, las variables influyentes y facilidad de implementación. Para ello, se utilizaron los mismos datos de entrenamiento (70%) y datos de prueba (30%). Los resultados, obtenidos con el software estadístico R Studio, indicaron que ambos modelos ofrecen buenas predicciones. La regresión logística mostró una precisión ligeramente superior con el 87.71%, en comparación con el 85.40% de los árboles de decisión. Además, el área bajo la curva -AUC-, para la regresión logística fue del 85%, mientras que para los árboles de decisión fue del 81%. La pérdida logarítmica también indicó diferencias, siendo del 31% para la regresión logística y del 38% para los árboles de decisión. Finalmente, los dos modelos proporcionaron resultados similares con una diferencia mínima basada en las variables explicativas utilizadas. La elección del modelo más adecuado en una entidad financiera dependerá de las variables que se consideran más relevantes para la predicción del riesgo crediticio y la facilidad de implementación con el sistema de información disponible en la institución.

Palabras clave: Árboles de decisión, Calificación del cliente, Comparación, Riesgo crediticio, Regresión logística.

ABSTRACT

This study looks at the credit rating of microcredits given by the Fernando Daquilema Ltda. Savings and Credit Cooperative in 2023. It does this by comparing the performance criteria of the regression model and decision trees in order to find the model that can predict things the most accurately. The research has a quantitative approach of a correlational and explanatory type. We compared the models using precision metrics, the confusion matrix, the ROC curve, influential variables, and ease of implementation. We used the same training data (70%) and test data (30%) for this comparison. The results, obtained with the statistical software R Studio, indicate that both models offer accurate predictions. Logistic regression showed a slightly higher accuracy of 87.71%, compared to 85.40% for decision trees. Additionally, the area under the curve (AUC) for logistic regression was 85%, while for decision trees it was 81%. The logarithmic loss also indicated differences, being 31% for logistic regression and 38% for decision trees. Finally, the two models provided similar results with a minimal difference based on the explanatory variables used. The best model for a financial institution will depend on the variables that are thought to be most useful for predicting credit risk and how easy it is to use with the institution's current information system.

Keywords: Decision trees, Customer rating, Comparison, Risk credit, Logistic Regression.

CAPÍTULO I

PROBLEMA

1.1. Planteamiento del problema

La crisis del COVID-19 provocó una transformación en la salud y estabilidad financiera en todo el mundo, el impacto más drástico se vio en los sistemas de ahorro y crédito de tipo cooperativo. Cuando una oferta laboral escasea y la producción se reduce, se crea una crisis del nivel más grave (Macías y Loor, 2022). Las limitaciones de movilidad impuestas por el gobierno para evitar la pérdida de vidas han tenido un efecto negativo también en la economía, generando un incremento de la morosidad por impago de créditos que se han concedido (Luque y Peñaherrera, 2021).

A nivel mundial, la gestión de riesgo crediticio es una preocupación central, especialmente en sectores microempresarios donde el acceso de crédito es vital para el desarrollo económico. Algunas investigaciones destacan el aumento de la morosidad como un desafío significativo que requiere la implementación de modelos estadísticos más precisos para evaluar el riesgo crediticio. Aunque la regresión logística ha sido un enfoque común, actualmente existe un crecimiento de interés en modelos alternativos como las técnicas de *machine learning* (Martínez Fernández, 2022).

En Ecuador, la situación económica de los últimos años ha tenido un impacto importante en los clientes con crédito a la hora de cancelar sus obligaciones. Es así como ciertas entidades comerciales implementaron planes de refinanciamiento para mantener sus niveles de morosidad dentro de parámetros estables (López, 2017). Las cooperativas de ahorro y crédito juegan un papel fundamental para el desarrollo de económico del país, enfrentan problemas relevantes, derivado de diversos factores como la pandemia, terremoto y apreciación de dólar (Congacha, 2023).

Debido a esto, resulta esencial adoptar modelos de evaluación crediticia más eficaces en el contexto de microcréditos (InSight, 2008). A pesar de que se han investigado varias técnicas, como el score crediticio, regresión logística y árboles de decisión en riesgo crediticio, hay una carencia de investigaciones que muestren de forma detallada la eficacia de la regresión logística en comparación con los árboles de decisión.

La cooperativa Fernando Daquilema Ltda. enfrenta problemas similares, ya que unos de los productos con un elevado índice de morosidad es la línea microcréditos (Pacific Credit Rating [PCR], 2023). La metodología actual utilizada por la institución para evaluar las solicitudes de crédito se basa principalmente en criterios subjetivos, sin aprovechar la base de datos la base de datos interna para incorporar variables relevantes ni utilizar ningún método estadístico (Congacha, 2023). Esto genera limitaciones en la medición de riesgo crediticio. Aunque se han estudiado enfoques para el control de riesgo de crédito, como la matriz de cosechas, indicadores financieros, modelo *logit*. No existen estudios que comparen a fondo la efectividad del modelo de regresión y árboles de decisión basados en precisión y facilidad de implementación.

La técnica más utilizada para predicción de una calificación crediticia en el ámbito financiero es análisis discriminante y regresión logística, siendo la más utilizada la regresión logística. Sin embargo, en la actualidad, los árboles de decisión han ganado la popularidad en la calificación crediticia y ahora se utilizan comúnmente para ajustar datos y predecir incumplimientos (Hernández, 2004).

Tanto la regresión como los árboles de decisión son métodos ampliamente empleados en diversas áreas de investigación, principalmente en la medición de riesgo crediticio. Sin embargo, cada uno tiene sus propias ventajas y desventajas, lo que lleva a distintos criterios de evaluación. Esta variedad de enfoques ha generado un debate sobre cuál método es más eficaz para estimar la calificación crediticia, lo que representa un desafío considerable. La controversia surge debido a las conclusiones contradictorias en los estudios; algunos indican que la regresión es superior, mientras que otros defienden los árboles de decisión. Este dilema no solo se centra en la precisión del modelo, sino también en su aplicación e interpretación.

Ante esta ausencia de un análisis comparativo que evalúa la efectividad de la regresión logística y los árboles de decisión en la estimación de la calificación crediticia genera incertidumbre en la gestión de riesgo crediticio. Por consiguiente, este estudio, se propone responder a la siguiente pregunta: ¿Cuál de los dos modelos entre regresión logística y árboles de decisión, es más eficaz en la estimación de la calificación crediticia en los microcréditos de la cooperativa Fernando Daquilema Ltda.?

1.2. Preguntas de investigación

La presente investigación formula las siguientes preguntas clave que guiarán el estudio:

¿Qué procesos realizar para transformar la cartera de microcréditos en una base de datos estructurada y útil para la aplicación de modelos de predicción?

¿Cuál es la eficacia y precisión de los modelos de regresión y árbol de decisión en la estimación de la calificación crediticia del microcrédito?

¿Qué métricas de desempeño se puede comparar entre el modelo de regresión y árboles de decisión en la determinación de la calificación crediticia?

1.3. Objetivos de investigación

1.3.1. Objetivo General

Evaluar la calificación crediticia de los microcréditos otorgados por la Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda. en el 2023, por medio de la comparación de criterios de desempeño entre el modelo de regresión y árboles de decisión.

1.3.2. Objetivos Específicos

1. Determinar las variables que influyen en la estimación de la calificación crediticia de los microcréditos de COAC Fernando Daquilema Ltda. en el 2023.

2. Estimar la calificación crediticia de los microcréditos de la COAC Fernando Daquilema Ltda. A través de la aplicación de los modelos de regresión y árboles de decisión.

3. Analizar la calificación crediticia de microcréditos otorgados en la COAC Fernando Daquilema Ltda. por medio de la comparación de la precisión predictiva y facilidad de implementación entre los modelos de regresión y árboles de decisión.

1.4. Justificación

La investigación compara modelos de regresión y árboles de decisión para evaluar la calificación crediticia en microcréditos en la Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda., abarcando el periodo comprendido entre enero y diciembre de 2023. Este marco temporal es fundamental para analizar la evolución y dinámica de la cartera de microcréditos, con un enfoque específico en esta línea crediticia. Se excluye el análisis de otras líneas de crédito, lo cual es necesario para profundizar en un segmento crítico que ha mostrado altos índices de morosidad, tanto en esta cooperativa como en otras entidades financieras.

La elección de la Cooperativa Fernando Daquilema para el desarrollo de esta investigación se fundamenta en su destacada trayectoria tanto a nivel local como nacional. La cooperativa no solo ofrece productos y servicios financieros a sectores vulnerables, sino que su cartera de microcréditos es un componente esencial de su operación. Este estudio es conveniente, ya que representa una oportunidad para aplicar los modelos propuestos en un entorno real, con el objetivo de identificar el modelo más efectivo para la evaluación de solicitantes de crédito.

La trascendencia social de este estudio radica en su capacidad para facilitar el acceso a microcréditos para emprendedores de bajos ingresos mediante métodos más eficientes y justos. Esto mejorará la inclusión financiera como acceso a servicios y productos financieros que ofrece la institución, esto generará oportunidades de empleo y desarrollo en los sectores rurales, impactando positivamente la economía local y nacional.

La investigación tendrá implicaciones prácticas significativas en el diseño de estrategias y programas para optimizar la evaluación crediticia. Esto permitirá desarrollar modelos específicos para la cooperativa Fernando Daquilema Ltda. y otras entidades financieras, facilitando una gestión más efectiva del riesgo crediticio.

Este estudio también ampliará el conocimiento actual de los métodos predictivos en microcréditos, tratando una carencia en el contexto local y nacional respecto a la comparación de modelos de regresión y árboles de decisión sobre cuál de los dos modelos es mejor en la estimación de la calificación crediticia y en base que métricas de precisión. Los resultados obtenidos podrían ser aplicables a principios más amplios, facilitando que otros investigadores examinen metodologías comparativas en diferentes contextos según los objetivos del estudio.

Desde una perspectiva metodológica, esta investigación se centra en el desarrollo y aplicación de modelos de regresión y árboles de decisión para estimar la calificación crediticia, lo que mejorará la comprensión de las relaciones entre variables cualitativas, cuantitativas y el riesgo crediticio. Este enfoque no solo aportará recomendaciones valiosas para futuras investigaciones en diversos contextos como la metodología de comparación y poblaciones, sino que también destaca la relevancia de la estadística en el desarrollo del modelo de predicción crediticia. Las herramientas analíticas y estadísticas facilitan la identificación de patrones y la evaluación de modelos de predicción, siendo esenciales para mitigar el riesgo de incumplimiento.

Por último, esta investigación se alinea con el Objetivo de Desarrollo Sostenible (ODS) 8 y la Línea de Investigación del Programa de Postgrado: "Aplicación de la Estadística en la solución de problemas del entorno", a la que se adscribe la investigación, su propósito es mejorar las condiciones de vida mediante un crecimiento económico que sea sostenido, inclusivo y sostenible. Además, de fomentar el empleo pleno, productivo y el trabajo decente para todos. También apoya el Plan Nacional de Desarrollo 2021-2025 al promover la inclusión financiera y el acceso a servicios económicos de alta calidad, destacando los objetivos 2 y 3. Esta alineación contribuye a un enfoque integral que integra el crecimiento económico, la inclusión social y la sostenibilidad. Como también el apoyo hacia el desarrollo de capacidades productivas y el emprendimiento que son fundamentales para la creación de empleo y la mejora de la calidad de vida en las zonas rurales.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

La fundamentación teórica de esta investigación se estructura en tres componentes esenciales. En primer lugar, la sección de antecedentes analiza investigaciones previas que han abordado la comparación de diferentes técnicas para la estimación de la calificación crediticia, destacando los hallazgos relevantes y las metodologías utilizadas.

A continuación, el marco teórico proporciona un aporte conceptual, delineando los principios y teorías que sustentan el análisis de los modelos de regresión y árboles de decisión como herramientas analíticas para la calificación crediticia, así como las interrelaciones entre las variables que afectan a los prestatarios de microcréditos.

Por último, el marco legal examina las normativas y regulaciones aplicables a la medición de la calificación crediticia, asegurando que la investigación sea sustentada bajo los parámetros legales.

2.1. Antecedentes de investigación

En cuanto a los estudios anteriores que respaldan esta investigación, es fundamental destacar aquellos que han utilizado tanto técnicas estadísticas convencionales como modelos de *machine learning* para comparar su efectividad en desarrollar un modelo óptimo de estimación de calificación crediticia. Estos estudios ofrecen una evaluación de las ventajas y limitaciones de cada método, proporcionando un contexto robusto para el análisis comparativo de esta investigación.

Curto (2023) llevó a cabo un estudio en España en colaboración con la Universidad de Valladolid. Comparó diversas técnicas de predicción de aprendizaje automático, se incluyen la regresión logística, el modelo probit, el árbol de decisión y la red neuronal. Los resultados demostraron que los modelos logit y probit ofrecen resultados prácticamente idénticos. Además, se evidenció que no hay diferencias significativas entre los resultados del modelo logit y su equivalente de aprendizaje automático, la regresión logística. Se han empleado varios criterios de evaluación, como la tasa de acierto promedio en la validación cruzada, la tasa de acierto global, la matriz de confusión y diversas métricas derivadas de esta última, como la sensibilidad, la especificidad, el valor F1, así como la curva ROC y su área bajo la curva (AUC).

Bennell *et al*. (2006) llevaron a cabo un estudio en Reino Unido en colaboración con escuela de negocios de *Aberystwyth*. Utilizaron un conjunto completo de datos de agencias de calificación y países durante el período 1989-1999, donde demostraron que las redes neuronales artificiales representaron una tecnología superior para calibrar y predecir calificaciones soberanas en relación con el modelado *probit* ordenado. Los resultados de las calificaciones crediticias soberanas presentados corroboran los hallazgos de otros investigadores de que las ANN son clasificadores altamente efectivos.

Ozturk *et al*. (2016) Exploraron el rendimiento de predicción de varias técnicas de inteligencia artificial para predecir calificaciones crediticias soberanas en una muestra heterogénea. Los resultados sugieren que los clasificadores de IA superan a la técnica estadística convencional en términos de predicción precisa. Según la medida de predicción precisa de uno y dos niveles, el rendimiento de predicción de los clasificadores de IA supera el 90% de precisión, mientras que el rendimiento del método estadístico convencional es de alrededor del 70%.

Sezgin (2006) enfocó en el proceso de calificación crediticia interno, utilizando datos reales de un banco turco sobre empresas manufactureras para construir varios modelos predictivos, incluyendo regresión logística, *probit*, análisis discriminante y árboles de clasificación y regresión. Se destaca la importancia de seleccionar una muestra óptima para mejorar el rendimiento de los modelos, así como la conversión de variables continuas en variables escaladas ordenadas para evitar problemas de escala. La evaluación del rendimiento de los modelos se realizó tanto dentro como fuera de la muestra, utilizando técnicas de validación sugeridas por el Comité de Basilea. Se concluye que, en la mayoría de los casos, el modelo de clasificación y árboles de regresión supera a las otras técnicas.

Caner (2023) evalúo el desempeño de modelos de riesgo crediticio estándar como regresión logística y *Probit*, junto con modelos avanzados como *Random Forest* y LightGBM. Se utilizaron métricas como la curva ROC, el valor de recuperación y la matriz de confusión para medir la efectividad de los modelos. Los resultados mostraron que la regresión logística superó al *probit* en la predicción del riesgo crediticio, obteniendo un AUC de 0,77 y una recuperación de 0,73, mientras que *Probit* tuvo un AUC de 0,67 y una recuperación de 0,42. *LightGBM* destacó con un AUC de 0,89 y una recuperación de 0,88, siendo superior al Random Forest en términos de recuperación.

Bensic *et al.* (2005) analizaron características relevantes para la calificación crediticia de préstamos a pequeñas empresas en un conjunto de datos con condiciones económicas específicas a través de métodos como regresión logística, redes neuronales y árboles de decisión CART. Compararon la precisión de los modelos extraídos, probando cuatro algoritmos de redes neuronales. Aunque no se encontraron diferencias estadísticamente significativas entre los modelos, el modelo probabilístico de redes neuronales obtuvo la mayor tasa de acierto y el menor error tipo I. Este modelo también mostró la mayor asociación con los datos y el menor costo total de clasificación errónea en todos los escenarios estudiados.

Manvitha y Rekha (2023) compararon el porcentaje la precisión de la humedad de las hojas utilizando modelos de regresión logística frente al modelo de árbol de decisión, donde el modelo de regresión obtuvo una precisión de 91.89% en comparación con la precisión de árbol de decisión con 80.24%.

Reddy y Karthikeyan (2022) desarrollaron un estudio en Pakistán en participación conjunta con conferencia internacional sobre matemática, ciencias actuariales, informática y estadística MACS. Evaluaron que tan bien es el algoritmo de árbol de decisión y regresión logística en la clasificación de fotografías de fuego y humo, utilizaron métricas como precisión, recuperación, puntuación f y número de precisión para evaluar el rendimiento del algoritmo. En comparación entre ambos modelos, el árbol de decisión funcionó mejor con 90.54% de precisión frente al modelo de regresión logística con 87.75%.

Boer *et al*. (2023) compararon algunos modelos de clasificadores entre ellas el modelo de árbol de decisión y regresión logística, en los resultados obtuvieron que el modelo de regresión logística mostró mejor precisión con su métrica en promedio de 92.61% frente a los demás modelos que resultaron con sus métricas por debajo de 90%.

Aji y Dhini (2019) señalaron que el score crediticio es una evaluación de la viabilidad de las solicitudes de crédito, en ese estudio utilizó el enfoque de minería de datos en donde se utilizó dos clasificadores las cuales son: árbol de decisión y *random forest*. El clasificador con mayor precisión es *random forest* con un 72.95%, en cambio, la peor precisión la realizó fue árbol de decisión con un 68.7%. La mejor sensibilidad realizada por *random forest* complementada con Ada-boost con 0.73. Se consideraron

como el mejor modelo en términos de prevenir el error tipo II que podría impactar en el aumento de la morosidad en una institución financiera.

Wang *et al*. (2022) predijeron el incumplimiento para préstamos personales de un experimento sobre un conjunto de datos que incluyó más de 20 mil préstamos personales en línea, las principales variables o características que incluyen en la industria de los prestatarios son: los ingresos, los antecedentes educativos, la edad, la ciudad, el género, los bienes raíces y el puntaje crediticio de Zhima. En este estudio utilizó 6 tipos de clasificadores como es: árbol de decisión, *random forest, Naive* Bayes, KNN, SVM, Neural Network, siendo la más afectiva la clasificación *Random Forest* con 94.5% de efectividad.

También existe estudios sobre la comparación de diferentes técnicas para la estimación de la calificación crediticia a nivel Latinoamérica. A continuación, menciona algunos autores que llevaron a cabo este estudio.

González (2023) mencionó que la forma efectiva de gestionar el riesgo crediticio es a través del uso de modelos estadísticos que, a partir de una muestra de clientes clasificados como buenos o malos, permiten predecir la probabilidad de incumplimiento, al utilizar modelos avanzados, las entidades financieras pueden tomar decisiones fundamentadas, establecer niveles de riesgo, asignar capital de manera eficiente y cumplir con los requisitos regulatorios.

Hernández (2004) indicó que la técnica más utilizada para predicción de una calificación crediticia en el ámbito financiero es análisis discriminante y regresión logística, siendo la más utilizada la regresión logística porque vincula la puntuación y la probabilidad de incumplimiento. Sin embargo, en la actualidad, los árboles de decisión han ganado la popularidad en la calificación crediticia y ahora se utilizan comúnmente para ajustar datos y predecir incumplimientos.

González (2023) realizó un estudio en México, en donde comparó el desempeño de dos algoritmos de aprendizaje supervisado, el árbol de decisión y la regresión logística, en la predicción de riesgo crediticio. Se evaluó la eficacia de los modelos de regresión logística frente a los árboles de decisión en la predicción de riesgo crediticio, encontrando que la regresión logística demostró una mayor eficacia con un 0.93 de exactitud, mientras

que los árboles de decisión obtuvieron una eficacia de 0.83 de exactitud al entrenar ambos modelos con el mismo número de muestras.

Martínez Fernández (2022) realizó la comparación de modelos de *Machine Learning* aplicado al riesgo de crédito a partir de las métricas que miden el rendimiento de los algoritmos que nos permiten calcular la probabilidad de incumplimiento de los clientes de la cartera, según los análisis realizados, la metodología más adecuados para predecir la probabilidad de incumplimiento corresponden a la regresión logística y Gradient Boosting, ya que sus métricas en los datos de muestras y validación fueron mejores que al promedio obtenido en los demás modelos.

Izquierdo Cruz (2019) comparó dos enfoques para evaluar el incumplimiento en una cartera de crédito especializada en financiamiento de motocicletas de una institución financiera del sector real. Por un lado, se empleó árboles de decisión como una técnica exploratoria de autoaprendizaje no paramétrica. Por otro lado, se utilizan modelos de elección discreta que se centran en el comportamiento individual al relacionar la decisión de elección con diversos factores. Ambos enfoques se enmarcan en modelos probabilísticos. Se observó que, en términos de tasas de clasificación correcta, los árboles de decisión presentan un rendimiento por debajo del 60%, mientras que los modelos discretos logran un 62.4% y 61.6% para los modelos logit y Probit respectivamente.

Rodríguez Avellaneda (2018) evalúo la efectividad de varios modelos como regresión probit, árboles de decisión, llegando a una conclusión de que el modelo más efectivo para la predicción de riesgo crediticio fue árboles de decisión, con un mínimo error de 11,3% el cual permitió contar con un ajuste adecuado, demostrando la efectividad de este tipo de modelos en el sector microcrédito.

Ossa y Jaramillo (2021) compararon la precisión de un modelo de regresión logística frente a algunos modelos de *Machine Learning* para la estimación del riesgo de crédito, como resultado obtuvo que el modelo más equilibrado al momento de evaluación fue el *Random Forest*, dado que fue el que presentó el mejor ajuste basándose en las métricas de exactitud evaluadas.

Gonzáles *et al*. (2022) plantearon la comparación de dos modelos para evaluación de riesgo crediticio para el préstamo de productos prepago. Los modelos presentados fueron árboles de decisión y regresión logística las cuales fueron comparados por medio

de la curva de ROC y el método de validación cruzada K-Fold. El mejor modelo para viabilidad de un préstamo de producto solicitado es el árbol de decisión, ya que al realizar una comparación con el modelo de regresión logística y su validación cruzada el indicador de exactitud fue del 86%.

Salazar Vergara (2021) indicó que la caracterización de la variable es muy importante, porque permite descomponer las variables que componen el problema de investigación. Las variables independientes que aportaría para explicar la variable dependiente se basan en de acuerdo con su significancia estadística para su contribución a la predicción del nivel de riesgo, las cuales son: tipo de crédito, nivel educativo, plazo del crédito, tasa de interés, calificaciones de riesgo, edad, género, ubicación geográfica, situación legal, empleo, ingresos, disponibilidad para embargo, historial de morosidad y saldo pendiente.

Támara *et al.* (2019) buscaron desarrollar un sistema de calificación crediticia que identifique las variables más influyentes en la toma de decisiones sobre la aprobación de créditos en entidades financieras específicas. Se emplearon dos técnicas estadísticas, la regresión logística y las redes neuronales, las cuales han sido ampliamente utilizadas en estudios de pronóstico recientes. El estudio analizó doce variables y creó dos modelos, dividiendo una base de 43,086 obligaciones en un 70% para entrenamiento y un 30% para verificación. La precisión del modelo fue del 71.65% en la predicción de clientes de manera exacta, con un aumento al 72.04% en la base de revisión, indicando una exactitud similar en ambas particiones.

Barreno (2012) desarrolló una investigación en Perú en junto con la Universidad de Lima. Demostró una metodología para la comparación de los modelos de clasificación de regresión logística y árboles de clasificación en el estudio de la deserción universitaria, en ambas metodologías usaron los mismos datos de entrada y los datos de prueba para su evaluación. Los resultados obtenidos fueron similares en ambos casos con 94% aproximadamente de clasificación correcta.

En ecuador también existe investigaciones que emplean métodos estadísticos y algoritmos de predicción para analizar la calificación crediticia en los socios que solicitan un préstamo.

Peréz *et al.*, (2022) desarrollaron un algoritmo de aprendizaje automático que predice la probabilidad de pago de un deudor y estabiliza una estrategia de cobranza de deudas. Se recopila, limpia y preprocesa un conjunto de datos no balanceados con 7.447.856 registros para entrenar un clasificador aleatorio de Forest, una máquina de refuerzo de gradiente, una regresión logística y un perceptrón multicapa utilizando una técnica de submuestreo aleatorio. El desempeño de los modelos se compara utilizando las métricas de evaluación de sensibilidad, especificidad y AUC. El algoritmo con mejor rendimiento es el Gradient Boosting Machine con una sensibilidad de 0,97, una especificidad de 0,93 y un AUC de 0,98 en el conjunto de validación.

Buitrón *et al.*, (2022) se entrenan modelos lineales para predecir la probabilidad de pago de un cliente a una agencia de cobranza de deuda ecuatoriana los primeros tres meses después de firmar un acuerdo de pago. Específicamente, se implementan una red neuronal de avance, regresión logística y modelos de conjunto de refuerzo de gradiente utilizando la metodología de minería de datos SEMMA, que comprende los pasos de muestrear, explorar, modificar, modelar y evaluar. Además, al analizar los resultados de los modelos, se identifican que las redes neuronales funcionan mejor que los modelos competidores en términos de precisión de clasificación.

Chiluiza *et al.*, (2023) llevaron a cabo un estudio en donde el objetivo consiste aplicar la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) y los modelos Random Forest (XGBoost), Regresión Logística y Redes Neuronales de aprendizaje supervisado, para implementar un modelo predictivo que permita evaluar el riesgo crediticio. Con el resultado de la aplicación del modelo predictivo, se concluye que el uso de herramientas de *Machine Learning* ayuda a optimizar la evaluación del riesgo crediticio en las entidades financieras. Una vez utilizada la metodología CRISP-DM para el análisis, desarrollo y evaluación de los modelos, se concluyó que el modelo más eficiente es el Random Forest.

2.2. Marco Teórico

Para garantizar su eficiente funcionamiento, es esencial que toda institución financiera enfoque sus esfuerzos en gestionar los diversos riesgos a los que se enfrenta como parte de su labor. Dado que la concesión de créditos representa la principal fuente de ingresos en el ámbito financiero, resulta fundamental que se realice una evaluación adecuada de los riesgos asociados a esta actividad, con el objetivo de maximizar su otorgamiento a individuos o empresas que ofrezcan garantías sólidas para minimizar la probabilidad de incumplimiento en la recuperación de los fondos prestados (Ayús *et al*., 2010).

2.2.1. Factores influyentes en la medición de la calificación crediticia en microcréditos

El proceso de identificación es el reconocimiento de todos los factores que, al presentar comportamientos adversos, originan un incremento en el riesgo de crédito, lo que significa, identificar tanto el riesgo potencial en la concesión de créditos nuevos, como el posible deterioro de la calidad crediticia de operaciones ya desembolsadas.

Actualmente, las entidades financieras se basan en los 5 c para evaluar las solicitudes de crédito, las cuales son: capacidad, carácter, colateral y condiciones (Vargas y Mostajo, 2014). Estas características de mediciones forman los diferentes factores o variables que influye en la concesión de un préstamo, y por consiguiente utilizando técnicas estadísticas evaluar la significancia de cada una de ellas. Aunque no existe un número óptimo de variables para la construcción de modelos estadísticos, ya que la selección de variables varía según el estudio y depende de la naturaleza de los datos y de factores culturales o económicos que pueden influir en la calidad del modelo (Abdou y Pointon, 2011).

Las variables cualitativas como estado civil, edad y el tipo de vivienda (propia, de familiares o alquilada), además de variables relacionadas intrínsecamente con la operación crediticia como el plazo, el número de créditos con la entidad financiera y el saldo deudor en el sistema financiero, son las que generan un modelo correctamente ajustado (Calixto y Casaverde, 2011). Los atributos como Capital financiero, nivel educativo, estado civil y la edad como datos básicos, sus operaciones y pagos generados, son factores que influye para la construcción de modelos de predicción (Borrero y Bedoya, 2020).

Reconocer el riesgo crediticio significa aceptar que cada transacción de crédito está acompañada de un grado de incertidumbre. Esto se debe a la presencia de varios factores que pueden influir en la capacidad y disposición del prestatario para cumplir con sus compromisos. Entre estos factores se encuentran la calidad crediticia del prestatario, su historial de pagos, su capacidad para generar ingresos suficientes que cubran sus deudas, la situación económica general, y otros riesgos específicos relacionados con el sector o industria en la que el prestatario opera (Pérez, 2017).

Las variables más importantes que se utilizan actualmente en la industria para predecir default son: Calificación crediticia del deudor, relación deuda-ingreso, relación préstamos-valor, historial de empleo, historial de pago, antigüedad del historial crediticio, tipo de crédito, movimiento de cuentas bancarias, Ahorros y certificados (García, 2024).

Abdou y Pointon (2011), algunas de las variables consideradas para un modelo de estimación de calificación crediticia incluyen: variables financieras como rentabilidad, liquidez y préstamos bancarios; información personal como edad, salario e historial bancario; historial crediticio, que abarca el tiempo de empleo y el tiempo con el banco; y variables de actividad económica.

2.2.2. Estimación de riesgo crediticio

El riesgo crediticio es una inquietud esencial en el sector financiero, en particular para las entidades que conceden préstamos. Este riesgo se relaciona con la posibilidad de que un prestatario incumpla sus compromisos de pago, lo cual podría ocasionar pérdidas económicas para el prestamista. Por consiguiente, es vital comprender y evaluar de manera adecuada el riesgo crediticio antes de extender crédito a cualquier individuo, empresa u otra entidad (Hernández, 2004).

Una sociedad financiera comercial es una entidad cuyo propósito principal es recolectar fondos a corto plazo para llevar a cabo operaciones de préstamo que faciliten la venta de bienes y servicios. Para las empresas que ofrecen servicios de crédito, es esencial contar con liquidez para cumplir con sus compromisos y mantener un flujo de efectivo constante que les permita seguir operando. Asimismo, estas empresas se enfocan en reducir el riesgo crediticio debido a la naturaleza de su actividad (Borrero y Bedoya, 2020).

La estimación precisa de riesgo de crédito es la clave para la rentabilidad de una organización. Si la entidad financiera no logra estimar de manera correcta un riesgo, sobrevalora los préstamos y pierde su participación de mercado o fija las tasas de interés demasiado bajo para cubrir las pérdidas esperadas lo que conduce a malos resultados o indicadores financieros, principalmente el índice de morosidad. Finalmente, el riesgo crediticio es una parte vital del valor actual neto (VAN) de los instrumentos financieros, podría incorporarse en varios sistemas de recomendación orientados al cliente y campañas de marketing que la organización podría implementar.

Para el proceso de clasificación de los clientes con determinadas características se utiliza los modelos de clasificación, dentro de ella se hace referencia a 2 enfoques las cuales son: Aprendizaje supervisado y aprendizaje no supervisado (Barreno Vereau, 2012). En este conjunto se encuentran varias variables independientes o explicativas $X_1, X_2, \dots, X_p$ medidas en n observaciones, junto con una variable de respuesta Y también medida en esas n observaciones. La naturaleza del aprendizaje supervisado radica en utilizar un conjunto de datos previamente etiquetados, donde cada observación tiene asignadas clases o etiquetas (Domínguez Martin, 2021).

- ✓ **Modelo Logit**

Desde sus inicios, la regresión logística ha sido ampliamente reconocida como un modelo de clasificación predictivo en estudios epidemiológicos y clínicos a lo largo de las últimas décadas. Actualmente, se utiliza de manera frecuente en diversos campos como la investigación biomédica, economía, finanzas, criminología, ingeniería, salud pública, política, biología de la vida silvestre y psicología. Sus aplicaciones van desde la evaluación de riesgos crediticios, la predicción de resultados en elecciones políticas, la asignación de becas, hasta la anticipación de casos de cáncer de próstata y la probabilidad de infección por VIH. (Vargas y Mostajo, 2014).

Para seleccionar el mejor subconjunto de variables es una tarea compleja, particularmente cuando se dispone de un gran número de variables predictoras y no se cuenta con información precisa sobre la relación exacta entre ellas. En ocasiones, la cantidad total de posibles modelos puede ser inmensa; por ejemplo, cuando hay más de k=20 variables predictoras, evaluar todas las combinaciones posibles de subconjuntos de variables puede resultar en un alto costo computacional. Por ende, las técnicas de

optimización combinatoria y las estrategias de selección de modelos son de gran importancia y resultan esenciales para explorar el espacio de soluciones (Acuña Collazos *et al.*, 2012).

Las estrategias más empleadas para seleccionar el mejor subconjunto de variables son los métodos "Stepwise". Este enfoque se basa en elegir el modelo óptimo de manera secuencial, añadiendo o eliminando una sola variable predictora en cada paso según ciertos criterios de evaluación. Los tres algoritmos comúnmente utilizados son: *Backward Elimination* (Eliminación hacia atrás), *Forward Selection* (Selección hacia adelante) y *Stepwise Selection* (Selección paso a paso). No obstante, estos métodos cuentan con una fundamentación teórica limitada en lo que respecta a la decisión sobre el orden en que las variables se incluyen o excluyen en la construcción del modelo estadístico (Acuña Collazos *et al.*, 2012).

La regresión logística o modelo *logit* se emplea cuando se requiere anticipar resultados binarios, como determinar si un correo es spam o no, o si una empresa quebrará o no, considerando que existen múltiples factores que pueden afectar estos resultados. Dado que solo hay dos posibles resultados, 0 o 1, se le denomina regresión binomial. Aquí, el valor de la variable dependiente representa una probabilidad que oscila entre 0 y 1, por lo que al redondearla se obtienen valores de 0 o 1 (Castaño y Ramírez, 2005). La regresión se fundamenta en la Ecuación 5.

Función Logística

(TAŞ, 2023) la expresión matemática del algoritmo comienza con una ecuación lineal, que es la razón principal por la que se encuentra la regresión en el nombre del algoritmo de regresión logística; sin embargo, el algoritmo agrega logaritmos a la ecuación y pasa la ecuación a través de la función sigmoide. Este método permite que el algoritmo convierta el resultado de una ecuación lineal en una probabilidad; por lo tanto, en problemas de clasificación binaria, el algoritmo estima las salidas de valor real en el rango de [0,1].

Para obtener un modelo de regresión logística a partir de modelos lineales, es necesario seguir una serie de pasos que se describen detalladamente en el siguiente apartado

Los pasos de operación matemática del algoritmo de regresión logística se pueden explicitar en las siguientes ecuaciones. Comienza con la representación específica de la fórmula del algoritmo de regresión logística como se muestra en la Ecuación 1.

$$P\left(\frac{D}{X_1}, X_2, X_3, \dots\ , X_n\right) = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n}\beta_i X_i\right)}}$$

Ecuación 1 *Algoritmo de Regresión Logística*

La fórmula que muestra en la Ecuación 2 corresponde a la función lineal de la regresión Logística.

$$y = \beta_0 + \sum_{i=1}^{n}\beta_i X_i$$

Ecuación 2 *Ecuación Lineal*

$$P = \beta_0 + \sum_{i=1}^{n}\beta_i X_i$$

Ecuación 3 *Función lineal de Regresión Logística*

En lugar de intentar sustituir directamente el valor de probabilidad por el valor Y, se intenta resolver este problema colocando el valor impar, como se visualiza en la Ecuación 4.

$$\frac{P}{1-P} = y = \beta_0 + \sum_{i=1}^{n}\beta_i X_i$$

Ecuación 4 *Función Logit*

El valor de probabilidad solo puede tomar valores de 0 a infinito positivo, y el valor y también puede tomar valores de menos infinito. En otras palabras, incluso si se sustituye el valor de probabilidad por el valor Y, el valor de probabilidad puede dar una salida de valor real restringida; por lo tanto, esta situación hará que el modelo funcione en un rango restringido de valores y los resultados del modelo no serán óptimos. Se toma el logaritmo del valor de la probabilidad, como se muestra en la Ecuación 5 y los valores restantes pueden ser cualquier valor entre menos infinito y más infinito para resolver el problema.

$$y = \log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{n}\beta_i X_i$$

Ecuación 5 *Logit o predictor lineal*

Se multiplica ambos lados de la ecuación por la exponencial tal como se visualiza en la Ecuación 6 y se resuelve para la probabilidad.

$$\exp\left[log\left(\frac{P}{1-P}\right)\right] = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)$$

Ecuación 6 *Predictor lineal, multiplicado por exponencial*

El resultado de la ecuación anterior se expresa en la Ecuación 7, conocido como Regresión logística.

$$\sigma(t) = P = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n} \beta_i X_i\right)}}$$

Ecuación 7 *Regresión Logística*

Donde t es $(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)$

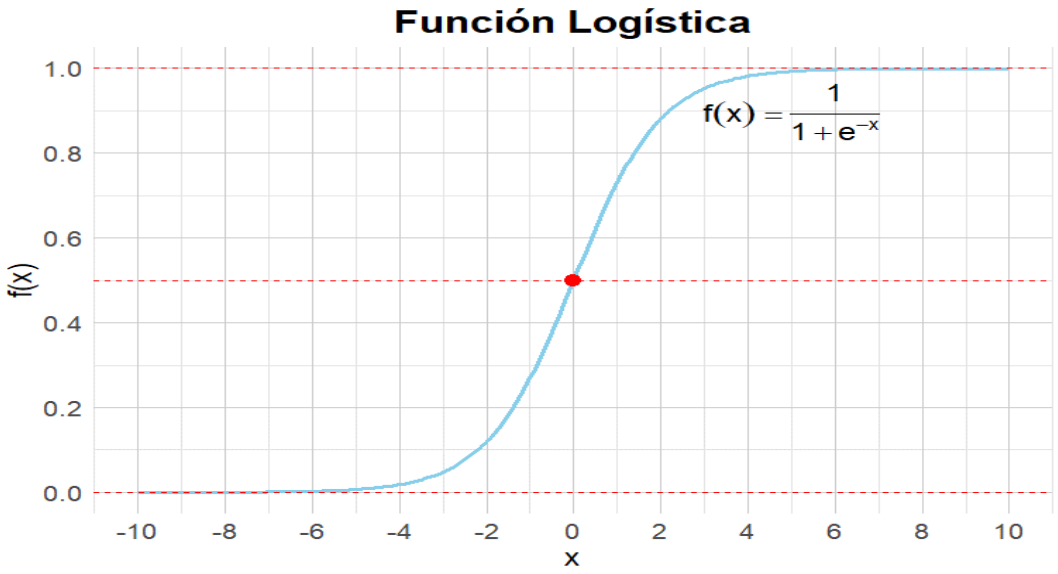
El resultado de la solución de la fórmula es la función logística, también llamada función sigmoide. Como muestra la ecuación anterior, la función sigmoide es una función de restricción utilizada en algoritmos de regresión logística. Esto se muestra en la Figura 1. Utilizando los datos de variables independientes para el entrenamiento del algoritmo, los valores producidos por el modelo se clasifican mediante la restricción de valores entre 0 y 1.

La función crea una curva en forma de S, lo que permite que el algoritmo de regresión logística resuelva problemas no lineales. El algoritmo de regresión logística con la función sigmoide clasifica los valores cercanos a 1 como positivos, mientras que los valores cercanos a 0 se clasifican como desfavorables (TAŞ, 2023).

El método del estimador de máxima verosimilitud se prefiere como función de costo, ya que el algoritmo de regresión logística se puede utilizar para problemas no lineales. La estimación de máxima verosimilitud estima los parámetros óptimos en los que opera un modelo en el proceso de entrenamiento.

Figura 1

Función Logística



Nota. Gráfica de la función logística. Fuente: Elaboración propia

Los valores de coeficiente determinados por el método de estimación de máxima verosimilitud permiten que el modelo maximice la probabilidad de observación. La verosimilitud es la medida de qué tan bien se ajusta el modelo al observar el conjunto de datos, y la estimación de máxima verosimilitud tiene como objetivo encontrar los valores de los coeficientes que mejor se ajusten a los datos. El algoritmo itera los valores de los coeficientes hasta que alcanza la máxima posibilidad en el proceso de optimización para encontrar esta máxima probabilidad (TAŞ, 2023).

Planteamiento del modelo

La variable respuesta Y es una variable dicotómica que toma valores 0 y 1,

$$Y = \begin{cases} 1\text{: Si se presenta el evento} & P(y_i = 1) = P_i \\ 0\text{: Si no se presenta el evento} & P(y_i = 0) = 1 - P_i \end{cases}$$

Para estimar los parámetros del modelo de regresión logística se emplea método de máxima verosimilitud. A diferencia de los modelos lineales estimados por mínimos cuadrados, los modelos de regresión logit son modelos no lineales, por lo tanto, se requiere procedimientos interactivos para su estimación (Barreno Vereau, 2012).

Interpretación del Modelo Logístico

Peña (2013) menciona que los parámetros del modelo son β_0, la ordenada en el origen, y $\beta_1 = (\beta_1, \dots , \beta_p)$, las pendientes. A veces también se utiliza como parámetros exp (β_0) y exp(β_i), que se denominan los odds ratios de probabilidad como se muestra en la Ecuación 8 y Ecuación 9, e indican cuanto se modifican las probabilidades por unidad de cambio en las variables x. En efecto, se deduce que:

$$\boldsymbol{Odds} = O_i = \frac{p_i}{1 - p_i} = \exp(\beta_0) * \prod_{j=1}^{p} exp\left(\beta_j\right)^{x_j}$$

Ecuación 8 *Razón entre probabilidades*

El cociente de las ratios de probabilidades (odds ratio) es:

$$\boldsymbol{odds\ ratio} = \frac{O_i}{O_k} = e^{\beta_K}$$

Ecuación 9 *Razón entre Odds*

Peña (2013) describe que el cociente indica cuanto se modifica la ratio de probabilidades cuando la variable X_h aumenta una unidad. Debido a que el coeficiente de cada variable asociado es igual al odds ratio, (Barreno Vereau, 2012) menciona que estos valores indican cuanto más probable es el éxito que el fracaso, dado un cambio unitario en la variable asociada. Por ende, se puede realizar las siguientes interpretaciones:

Si odds ratio $(X_h) < 1$ indica que el evento de interés disminuirá por cada incremento unitario de la variable X_h, manteniendo constantes las otras variables (Barreno Vereau, 2012).

Si odds ratio $(X_h) = 1$ indica que el evento de interés mantendrá sin variación por cada incremento unitario de la variable X_h, manteniendo constantes las otras variables. En este caso, el evento de interés, que es riesgo se mantendrá constante en cada incremento unitario de la variable X_h (Barreno Vereau, 2012).

Si odds ratio $(X_h) > 1$ indica que el evento de interés, probabilidad de que un cliente tenga un riesgo alto, aumentará por cada incremento unitario de la variable X_h, manteniendo constantes las otras variables (Barreno Vereau, 2012).

Prueba de hipótesis para los coeficientes del modelo

Poma Salcedo (2002), generalmente la estimación del modelo de regresión logística, al igual que en el modelo de regresión lineal múltiple, se realiza con diferentes propósitos, los cuales son: Evaluar si el coeficiente de una variable explicativa es igual a cero, Verificar si los coeficientes de un grupo de variables explicativas iguales a cero, determinar la calidad del ajuste global del modelo.

Prueba de Wald

Esta prueba se usa para evaluar la significancia estadística de cada variable explicativa o regresor, el estadístico de prueba para n pequeño y n grande se visualiza en Ecuación 10, Ecuación 11 y la hipótesis planteada es la siguiente:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

El estadístico de prueba:

$$W = \frac{(\hat{\beta}_j)^2}{\hat{\sigma}^2(\hat{\beta}_j)}$$

Ecuación 10 *Estadístico de Wald*

Bajo H_0, $W \sim \chi^2_{(1)}$ y para n grande se aplica la siguiente formula:

$$z = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \sim N\left(\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}, 1\right)$$

Ecuación 11 *Estadístico Wald para n grande*

Entonces z^2 se distribuye como una $\chi^2_{(\xi,1)}$ con parámetros de no centralización. Si la variable explicativa es categórica, los grados de libertad es igual al número de categorías menos uno.

Prueba Chi-Cuadrado

(Poma Salcedo, 2002) indica que esta prueba sirve para verificar si los coeficientes de un grupo de variables explicativas son iguales a cero, el estadístico de esta prueba se visualiza en Ecuación 12 y la hipótesis estadística planteados son:

$$H_0: \beta_1, = \beta_2 = \cdots = \beta_q = 0$$

$$H_1: \beta_j \neq 0, Al\ menos\ un\ j = 1,2, \ldots., q$$

La estadística de prueba para esta hipótesis es la siguiente:

$$\chi_q^2 = -2\left[\ln\left(L_{p-q}\right) - Ln\left(L_p\right)\right]$$

Ecuación 12 *Estadístico Chi- cuadrado*

Bajo H_0 la estadística planteada sigue una distribución Chi cuadrada con q grados de libertad. Esta estadística también se usa para verificar si una variable independiente muestra una asociación significativa con la variable respuesta ante la presencia de otras variables.

Prueba Chi-Cuadrado de Pearson

Esta estadística sirve para determinar la calidad del ajuste global del modelo de regresión la cual se visualiza en Ecuación 13. Esto se basa en la comparación de los valores reales y_i con sus respectivas probabilidades predichas, π_i

Las hipótesis estadísticas definidas para esta prueba es la siguiente:

$$H_0: \beta_0, = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_j \neq 0, Al\ menos\ un\ j = 1,2, \ldots., q$$

La prueba se basa en la estadística de Chi-cuadrado de Pearson. Definida por:

$$\chi^2 = \sum_{i=1}^{n} \frac{r_i^2}{v_{ii}}$$

Ecuación 13 *Chi-cuadrado de Pearson*

Donde:

$$r_i = (y_i - \hat{\pi}_i)$$

$$v_{ii} = Diag\left(\hat{V}\right) = \hat{\pi}_i(1 - \hat{\pi}_i)$$

Bajo H_0 la estadística $\chi^2 = \sum_{i=1}^{n} \frac{r_i^2}{v_{ii}}$ tiene una distribución asintótica Chi-cuadrado con $\left(n - (k+1)\right)$ grados de libertad.

Desvianza

Otra manera de probar el ajuste global del modelo es mediante la estadística llamada Desvianza, es análogo a la suma de cuadrados de los residuales del modelo de regresión lineal múltiple que se aprecia en la Ecuación 14.

El planteamiento de hipótesis es la siguiente:

$$H_0: \beta_1, = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_j \neq 0, Al\ menos\ un\ j = 1,2, \ldots., k$$

Se usa para evitar la inestabilidad de estadística Chi-cuadrado de Pearson y la desvianza está dado por:

$$D_p = \sum_{i=1}^{n} d_i^2$$

Ecuación 14 *Estadística Desvianza*

Donde:

$$d_i \begin{cases} \sqrt{-2\log(\hat{p}_i)} & si\ y_i = 1 \\ \sqrt{-2\log(1-\hat{p}_i)} & si\ y_i = 0 \end{cases} ; j = 1,2, \ldots, n$$

Bajo H_0 la Desvianza es la misma que la distribución Chi-cuadrado de Pearson. Es decir, se distribuye $\chi^2_{(n-(k+1))}$ y mide la discrepancia entre el modelo bajo investigación y el modelo saturado.

La estadística $D_p = \sum_{i=1}^{n} d_i^2$ para el modelo de regresión logística está dada por la Ecuación 15.

$$D = -2\sum(y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i))$$

Ecuación 15 *Desvianza para Regresión Logística*

Cuando el modelo bajo investigación es verdadero se compara el valor D con el valor crítico χ^2_{n-p} de una distribución χ^2 a un nivel de significancia igual a $\propto$, por lo tanto: $Si\ D > \chi^2_{n-p}$ el modelo se rechaza y $Si\ D \leq \chi^2_{n-p}$ el modelo no se rechaza. Donde $p = k + 1$.

La prueba de Wald se utiliza para determinar la significancia estadística de cada variable explicativa en un modelo. Por otro lado, la prueba Chi-cuadrado evalúa la significancia de los coeficientes en un modelo logístico, verificando si las variables explicativas mejoran el ajuste. La estadística Chi-cuadrado de Pearson se emplea para evaluar la calidad del ajuste general del modelo, al comparar los valores observados con las probabilidades estimadas. De manera similar, la Desviancia también mide el ajuste global y es comparable a la suma de cuadrados de los residuos en la regresión lineal múltiple, permitiendo evaluar qué tan bien el modelo se adapta a los datos observados (Poma Salcedo, 2002).

- ✓ **Modelo de árbol de decisión binario**

Los clasificadores con estructura de árbol binario se desarrollan mediante una serie de divisiones recursivas del conjunto inicial, denominado X, en dos subconjuntos descendentes. Como se muestra en la Figura 2**¡Error! No se encuentra el origen de la referencia.**, los subconjuntos X2 y X3 son disjuntos, es decir, $X = X_2 U X_3$. De manera similar, los conjuntos X4 y X5 también son disjuntos, ya que $X_4 \cup X_5 = X_2$. Los subconjuntos que no se subdividen nuevamente se denominan conjuntos terminales. Se representa los subconjuntos terminales con un rectángulo y los no terminales (o de decisión) con un círculo. Los subconjuntos terminales constituyen una partición de X. Los subconjuntos no terminales se generan a partir de divisiones basadas en condiciones aplicadas a alguna de las coordenadas $x = (x_1, x_2, \dots, x_p)$ (Arana, 2021).

Los árboles de decisión pertenecen a la clase de aprendizaje automático supervisado, utilizan con frecuencia porque son fáciles de implementar, pueden interpretarse fácilmente, se aplican a variables cualitativas, cuantitativas, continuas y discretas y brindan resultados confiables (Kaur y Wasan, 2006).

Un árbol de decisión es un modelo de predicción que utiliza reglas binarias (sí/no) para dividir las observaciones según sus atributos, lo que permite prever el valor de una variable de respuesta (Amat, 2020). Los métodos basados en árboles han ganado importancia en el área de la predicción, ya que proporcionan resultados sólidos para una amplia gama de problemas. Este documento examina cómo se construyen y aplican los árboles de decisión para clasificación, los cuales son elementos clave en modelos predictivos más avanzados, como los bosques aleatorios y las máquinas de aumento de gradiente (Villalba et al., 2023).

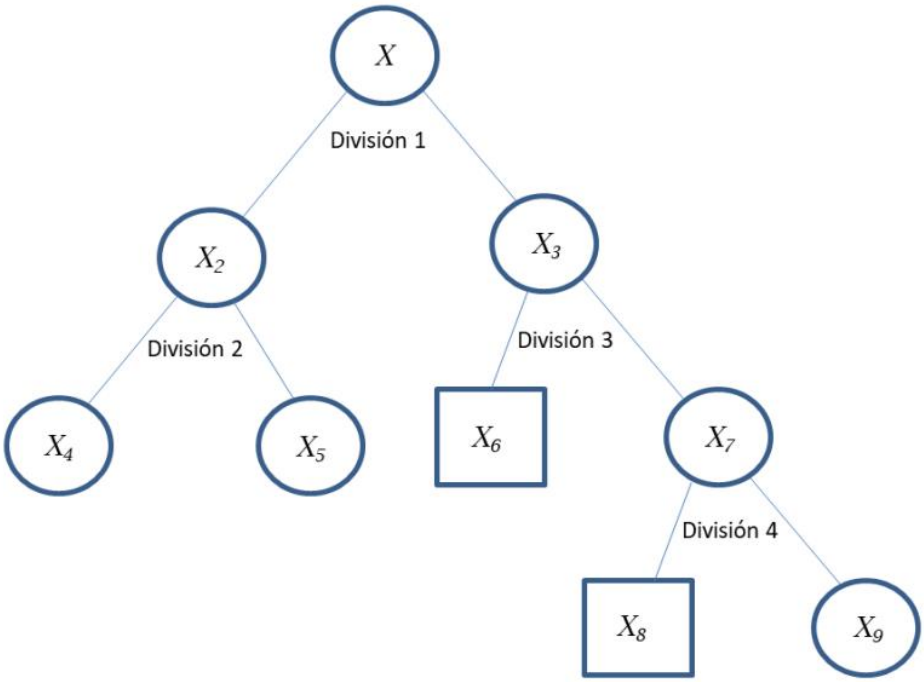
Las ventajas de la construcción del modelo de árboles de decisión son la simplicidad, potencia y estabilidad. El primer factor consiste en la comprensión ya sea de persona natural o a la vez una institución, de cómo funciona y que es lo que predice, por su parte la potencia nos da entender una medida en su capacidad de discriminar correctamente los solicitantes de un préstamo en clientes buenos y malos; y por último la estabilidad hace referencia a la duración en el tiempo sobre su capacidad de discriminación y de esa manera pueda identificar los cambios en la calidad de cartera bruta frente a la cartera improductiva de la entidad financiera (Arana, 2021).

Villalba *et al.* (2023) los árboles de decisión son modelos predictivos fáciles de interpretar, incluso cuando las relaciones entre los predictores son complejas, y pueden manejar tanto predictores numéricos como categóricos, lo cual depende de la biblioteca utilizada, sin necesidad de variables ficticias. Aunque los modelos basados en un solo árbol pueden visualizarse incluso con un gran número de predictores, su capacidad interpretativa es especialmente útil para la exploración de datos, permitiendo identificar automáticamente los predictores más relevantes de forma rápida y eficiente. Además, los árboles de decisión son aplicables tanto a problemas de regresión como de clasificación. Sin embargo, presentan ciertas limitaciones: son sensibles a datos de entrenamiento desequilibrados, perdiendo precisión cuando una clase domina sobre la otra; también, al trabajar con predictores continuos, pueden perder parte de la información al clasificar los valores en divisiones de nodos, y no pueden extrapolar más allá del rango de valores observados en los datos de entrenamiento.

Bae *et al*. (2021) recomienda el uso de algoritmos de balanceo de datos como SMOTE, ROSE y ADASYN. Dada la naturaleza particular de los datos y el objetivo del estudio, se ha elegido el algoritmo ROSE. Esto se debe a que el paquete ROSE proporciona funciones especialmente diseñadas para abordar problemas de clasificación binaria en contextos de clases desequilibradas (Lunardon *et al.*, 2014).

Figura 2

Estructura de árbol de decisión



Nota. Estructura de árbol de decisión obtenida. Fuente: (Arana, 2021)

El clasificador basado en árbol binario predice la clase para un caso por ejemplo para predecir si un cliente que solicita un préstamo presenta características de riesgo alto o riesgo bajo, a partir de las condiciones de la primera división, se determina si x se asigna al subconjunto X2 o X3. Si x se asigna al subconjunto X3, se someterá a las condiciones de la tercera división para decidir si ahora pertenece al subconjunto X6 o X7. Si x continúa a través de las divisiones sucesivas y sus respectivas condiciones, inevitablemente terminará en un subconjunto terminal como se muestra en la Figura 2. El proceso completo de construcción de un clasificador de árbol binario consta de tres elementos:

1. La selección de condiciones asociadas a cada división a partir de un conjunto de posibles condiciones.
2. La decisión sobre cuándo seguir dividiendo cada nodo o convertirlo en un nodo terminal.
3. La asignación de una clase específica a cada nodo terminal.

Dado que los datos proporcionados para el presente estudio son de tipo cualitativo categórico, como género, tipo de producto y estado de crédito, el modelo de árboles de decisión binaria se presenta como el método adecuado para los objetivos de la investigación, ya que permite construir un modelo adecuado y fácil de interpretar

2.2.3. Análisis de comparación entre el modelo de regresión y árboles de decisión.

Para el análisis comparativo, se emplea una metodología uniforme para aplicar cada modelo de clasificación evaluado. Las etapas que se compara incluyen la preparación de los datos de entrada, donde se preparan y transforman los datos para asegurar su calidad y compatibilidad.

La evaluación del modelo, que consiste en medir la significancia de las variables y modelo global; la interpretación del modelo, en la cual se identifican y analizan las variables más influyentes en las predicciones

Por último, la prueba del modelo para comparar los dos modelos se utiliza las siguientes métricas de precisión: como la matriz de confusión, exactitud del modelo, tasa de error, sensibilidad, especificidad, tasa de falsos positivos, tasa de falsos negativos, pérdida logarítmica, curva ROC, el área bajo la curva (AUC) para evaluar su robustez y consistencia en diversos escenarios (Barreno, 2012).

Barreno, (2012), la preparación de datos es esencial para desarrollar modelos de regresión y árboles de decisión. En modelos de regresión, es crucial transformar variables cualitativas para que sean interpretables por el algoritmo. La conversión de variables con dos categorías es sencilla, pero las variables *dummy* con más categorías presentan desafíos interpretativos. Los árboles de decisión no requieren una transformación exhaustiva, aunque esta puede mejorar la precisión de los resultados.

La evaluación del modelo para la regresión se realiza mediante análisis individual de variables y el modelo completo, utilizando pruebas estadísticas y la matriz de confusión. En contraste, los árboles de decisión se evalúan principalmente con la matriz de confusión, lo que limita las opciones estadísticas.

La interpretación de resultados varía entre ambos métodos. La regresión logística, basada en una función exponencial, es menos intuitiva que la regresión lineal, requiriendo reemplazo de valores en la ecuación final. En cambio, los árboles de decisión ofrecen una interpretación más accesible mediante reglas de decisión claras.

La prueba del modelo es crucial, evaluando resultados con datos de prueba, la regresión logística evalúa resultados mediante matriz de confusión, mientras que árboles de decisión también realiza mediante esta métrica.

Matriz de confusión

Es una matriz de dimensión 2x2 que compara las predicciones del modelo frente la clase real a la que pertenecen los individuos de los datos de la prueba tal como se visualiza en la tabla 1.

Tabla 1

Matriz de confusión

Valor	**Valor predicho**	
Real	Positivo	Negativo
Positivo	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Negativo	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Nota. Estructura de matriz de confusión (Martínez Fernández, 2022)

Cada columna de la matriz representa el número de predicciones obtenidas por el modelo, mientras que cada fila representa la clase real como se muestra en la Tabla 1. En la diagonal principal se pueden observar los individuos clasificados correctamente (Verdaderos positivos y negativos), mientras que la diagonal secundaria nos indica los errores de la clasificación (falsos positivos y negativos) (Martínez Fernández, 2022).

Verdaderos positivos (VP): Cantidad de observaciones clasificados correctamente.

Verdaderos negativos (VN): Cantidad de observaciones clasificados correctamente como negativos.

Falsos positivos (FP): Cantidad de observaciones clasificados incorrectamente como positivos.

Falsos negativos (FN): Cantidad de observaciones clasificados incorrectamente como negativos.

Métricas de evaluación

Las métricas para evaluar el rendimiento y la calidad de los modelos derivan de los valores de la matriz de confusión las cuales son:

Exactitud: Proporción de predicciones correctas

$$Exactitud = \frac{VP + VN}{Total} = \frac{VP + VN}{VP + FP + FN + VN}$$

Ecuación 16 *Exactitud del modelo*

En el caso de tener clases desbalanceados, si el modelo predice siempre la clase mayoritaria, siempre tendría un excelente nivel de Accuracy, esta situación ocurre en el nivel de un análisis de datos descriptivo y puede confundir a un observador ingenuo; por eso es necesario analizar el nivel de Accuracy que se muestra en la Ecuación 16 de la mano de otros indicadores Ossa y Jaramillo (2021).

Tasa de Error: Proporción de observaciones clasificadas incorrectamente.

Ecuación 17 *Tasa de Error de clasificación*

Sensibilidad: Conocido también como tasa de verdaderos positivos, es la proporción de casos positivos que fueron correctamente identificados

$$Sensibilidad = \frac{VP}{Total\ positivos} = \frac{VP}{VP + FN}$$

Ecuación 18 *Tasa de verdaderos positivos*

Especificidad: Conocido también como tasa de verdaderos negativos, es la proporción de casos negativos correctamente identificados.

$$Especificidad = \frac{VN}{Total\ negativos} = \frac{VN}{VN + FP}$$

Ecuación 19 *Tasa de verdaderos negativos*

Tasa de falsos positivos: Probabilidad de que se dé un resultado positivo cuando el valor verdadero es negativo:

$$TFP = 1 - especificidad = \frac{FP}{Total\ negativos} = \frac{FP}{VN + FP}$$

Ecuación 20 *Tasa de falsos positivos*

Tasa de falsos negativos: Probabilidad de que la prueba pase por alto un verdadero positivo, es decir, que se dé un resultado negativo cuando el verdadero valor es positivo.

$$TFN = 1 - sensibilidad = \frac{FN}{Total\ positivos} = \frac{FN}{VP+FN}$$

Ecuación 21 *Tasa de falsos negativos*

Pérdida logarítmica

El log-loss mide la discrepancia entre las probabilidades predichas por un modelo y las etiquetas verdaderas de los datos. Un valor de log-loss más bajo indica mejores predicciones (Ámbar, 2023).

El log-loss toma como entrada un vector de etiquetas verdaderas y una matriz de probabilidades de predicción. Para un conjunto de K clases y un conjunto de N puntos de datos, el log-loss se calcula utilizando la Ecuación 22.

$$LLOSS(\mu;\sigma) = -\frac{1}{N}\sum_{i=1}^{N}\ \sum_{k=1}^{K}\left([\sigma_i = k] * Ln(u_{i,k})\right)$$

Ecuación 22 *Pérdida Logarítmica*

Donde:

$$[\sigma_i = k] = 1\ si\ \sigma_i = k\ y\ 0, de\ lo\ contrario$$

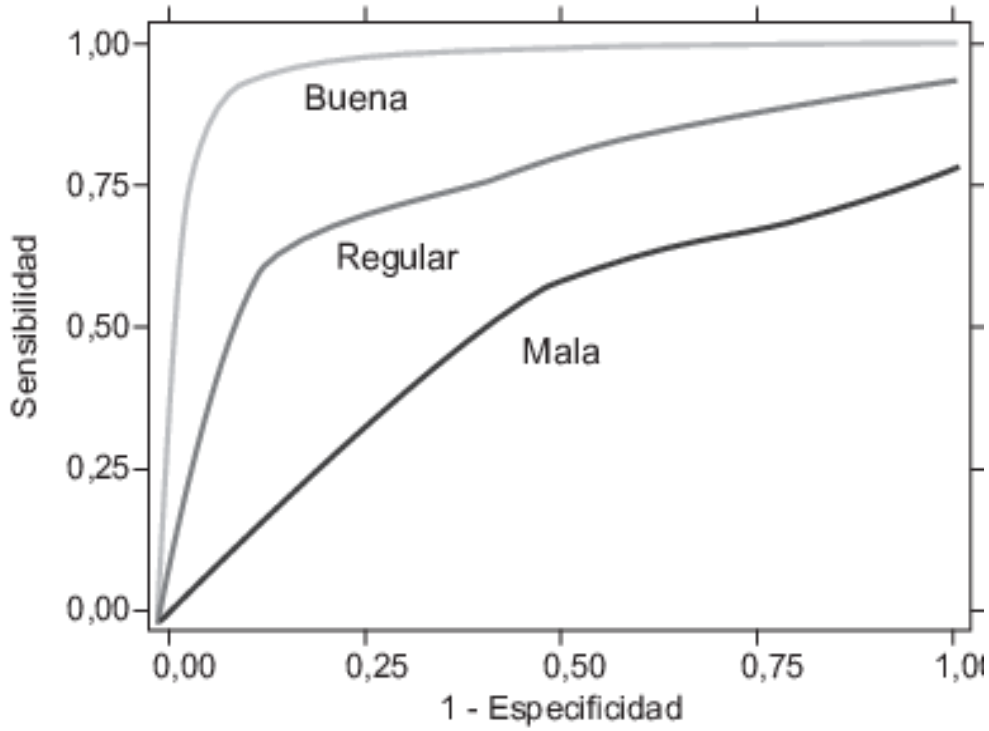
Curva de ROC

Al realizar una clasificación de individuos basada en la probabilidad de pertenencia a una clase, es esencial definir un umbral de decisión conocido como punto de corte. Este punto determina el nivel de probabilidad a partir del cual un individuo se asigna a una clase específica. Por lo general, el valor predeterminado de la mayoría de los algoritmos de clasificación es 0.5, lo que implica que si la probabilidad estimada es igual o mayor a 0.5, el individuo se clasificará en la clase 1, y si es menor a 0.5, se clasificará en la clase 0. Por consiguiente, la predicción realizada por el algoritmo de clasificación está sujeta al valor del punto de corte, lo que resulta en diferentes matrices de confusión y métricas para cada valor posible (Martínez Fernández, 2022).

La curva ROC (Receirver Operating Characterristic) es una representación gráfica, como se visualiza en la Figura 3, donde el eje Y corresponde a la tasa de verdaderos positivos (sensibilidad) y el eje X corresponde a la tasa de falsos positivos (1 – especificidad) de cada uno de los posibles puntos de corte (Martínez Fernández, 2022).

Figura 3

Ejemplo de Curva ROC



Nota. Curva Roc. Fuente: (Martínez Fernández, 2022).

La manera de cuantificar el rendimiento de curvas de ROC es a través del área bajo la curva (AUC, área under the curve):

$$\text{AUC} = \frac{\text{sensibilidad} - (1 - \text{especificidad}) + 1}{2}$$

Ecuación 23 *Área bajo la curva -AUC*

Este valor indica la capacidad del modelo para diferenciar entre las clases. Es importante destacar que un clasificador perfecto tendría un área de 1, mientras que un clasificador sin capacidad predictiva tendría un área de 0,5. A medida que el AUC se aproxima a 1, mayor será la capacidad del modelo para discriminar entre las clases como riesgo alto y riesgo bajo. En cambio, si la curva ROC coincide con la diagonal de referencia, el modelo se considera no discriminativo (Martínez Fernández, 2022).

a) Índice de Youden

$$J = Sensibilidad + Especificidad - 1$$

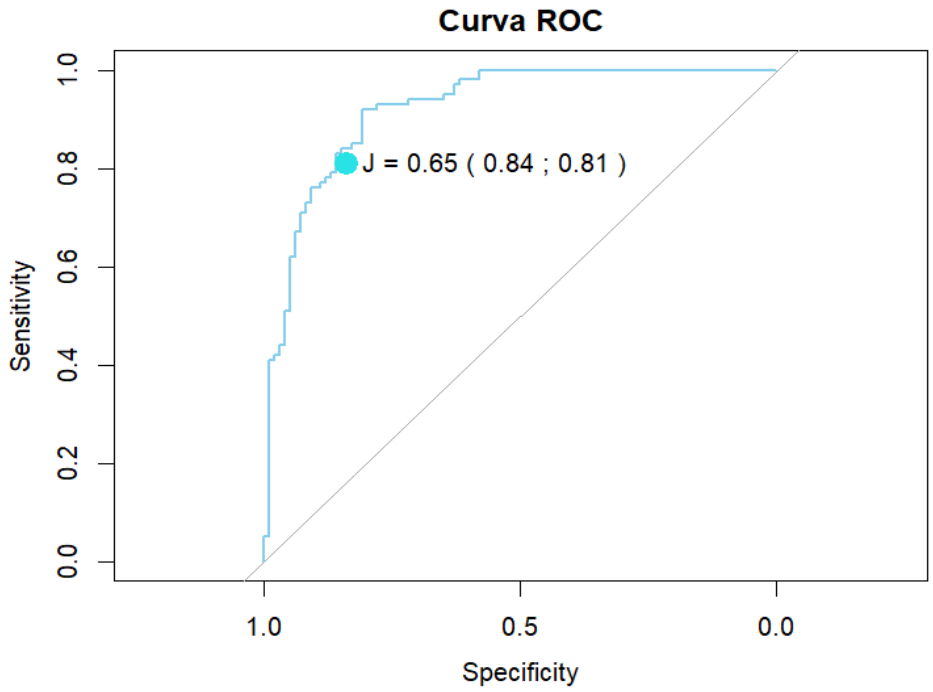
Ecuación 24 *Índice de Youden*

Los posibles valores del índice Youden obtenido mediante Ecuación 24 se encuentran en el rango $[0,1]$, donde valores próximos a 0 indican que el modelo tiene una baja capacidad discriminatoria, mientras que valores cercanos a 1 sugieren que no hay falsos positivos y que todas las observaciones se clasifican correctamente. Por ende, para tener un buen modelo es deseable alcanzar un valor elevado del índice de Youden (Martínez Fernández, 2022).

En la Figura 4 se presenta un ejemplo con el punto de corte óptimo determinado mediante el índice de Youden, que es 0.65. Clasificando según este punto de corte, se logra la mayor sensibilidad y especificidad para el modelo, con valores de 0.84 y 0.81(Martínez Fernández, 2022).

Figura 4

Ejemplo de Punto de corte Óptimo



Nota. Curva de ROC, punto de corte óptimo: 0.65, especificidad: 0.81 y sensibilidad: 0.84. Fuente: Elaboración propia

2.1. Marco Legal

En el artículo 310 de la constitución establece que el sector financiero público tendrá como finalidad la prestación sustentable, eficiente, accesible y equitativa de servicios financieros. El crédito que otorgue se orientara de manera preferente a incrementar la productividad y competitividad de los sectores productivos que permitan alcanzar los objetivos del Plan de Desarrollo y de los grupos menos favorecidos, a fin de impulsar su inclusión activa en la economía (Banco Central del Ecuador [BCE], 2020).

De la misma manera en el artículo 311 de la constitución indica que el sector financiero popular y solidario se compondrá de cooperativas de ahorro y crédito, entidades asociativas o solidarias, cajas y bancos comunales, cajas de ahorro. Las iniciativas de servicios del sector financiero popular y solidario y de las micro, pequeñas y medianas unidades productivas, recibirán un tratamiento diferenciado y preferencial del estado, en la medida en que impulsen el desarrollo de la economía popular y solidaria (BCE,2020).

Superintendencia de la Economía Popular y Solidario (SEPS, 2023) en su artículo N.° 5 establece que los elementos generales que deben tomarse en cuenta para calificar a los activos de riesgo en las distintas categorías. A efectos de identificar el perfil de riesgo de los sujetos de crédito comercial a continuación se describe las características de los factores de riesgo para cada una de las nueve categorías. Sin embargo, estas características no son determinantes para clasificar a un sujeto de crédito en una u otra categoría de riesgo, ya que en el análisis en conjunto de los factores serán los que determinen la calificación

De acuerdo con la normativa de control para la gestión del riesgo de crédito y la constitución de provisiones en fundaciones y corporaciones civiles cuyo principal objetivo es otorgar créditos, en conformidad con el artículo 8 del reglamento general de la ley orgánica de apoyo humanitario para enfrentar la crisis sanitaria generada por el COVID-19, de la RESOLUCIÓN Nro. SEPS-IGT-IGS-INR-INTIC-INGINT-0293 en la SUBSECCIÓN IV: DE LA CALIFICACIÓN DE RIESGOS:

Artículo 14.- Criterios de calificación. - Las entidades están obligadas a clasificar su cartera de crédito basándose en la morosidad y el segmento de crédito correspondiente, siguiendo los criterios establecidos en la resolución mencionada SEPS (2023).

Artículo 15.- Segmentos de crédito y tasas de interés activas. - Para segmentar la cartera de crédito, las entidades deben acatar las disposiciones del Capítulo IX "Normas que regulan la segmentación de la cartera de crédito de las entidades del Sistema Financiero Nacional" del Título II "Sistema Financiero Nacional", del Libro I "Sistema Monetario y Financiero", de la Codificación de Resoluciones Monetarias, Financieras, de Valores y Seguros. La definición de montos y plazos se realizará considerando las condiciones y la capacidad de pago de cada sujeto de crédito, así como la tecnología crediticia que desarrolle cada entidad para tal fin. Adicionalmente, las entidades deben observar y considerar las disposiciones relacionadas con las tasas de interés establecidas por la autoridad competente (BCE,2020).

CAPÍTULO III

METODOLOGÍA

3.1. Descripción del área de estudio

El estudio se lleva a cabo en la Cooperativa de Ahorro y Crédito (COAC) “Fernando Daquilema”, situada en Riobamba, provincia de Chimborazo, en la región central de Ecuador como se muestra en la Figura 5, con 19 sucursales a nivel nacional. Esta entidad legalmente constituida en el país se dedica a actividades de intermediación financiera y responsabilidad social con más de 159.000 socios en todo territorio nacional, operando bajo la supervisión de la Superintendencia de Economía Popular y Solidaria.

La cooperativa se rige por las leyes y regulaciones establecidas en la Ley Orgánica de la Economía Popular y Solidaria y del Sector Financiero Popular y Solidario, así como por su Reglamento General y las Resoluciones de la Superintendencia de Economía Popular y Solidaria.

Esta organización fue establecida el 25 de julio de 2005 en la parroquia de Cacha. Bajo el liderazgo de sus nuevos ejecutivos y dirigentes, se propuso convertirse en un vehículo financiero para la comunidad Puruwá, la confianza de sus socios se basa en el apego a los valores cristianos y culturales originarios. Estos principios han permitido que la institución sea una herramienta fundamental para que numerosos socios eviten caer en manos de prestamistas ilegales, además de fomentar una cultura de ahorro.

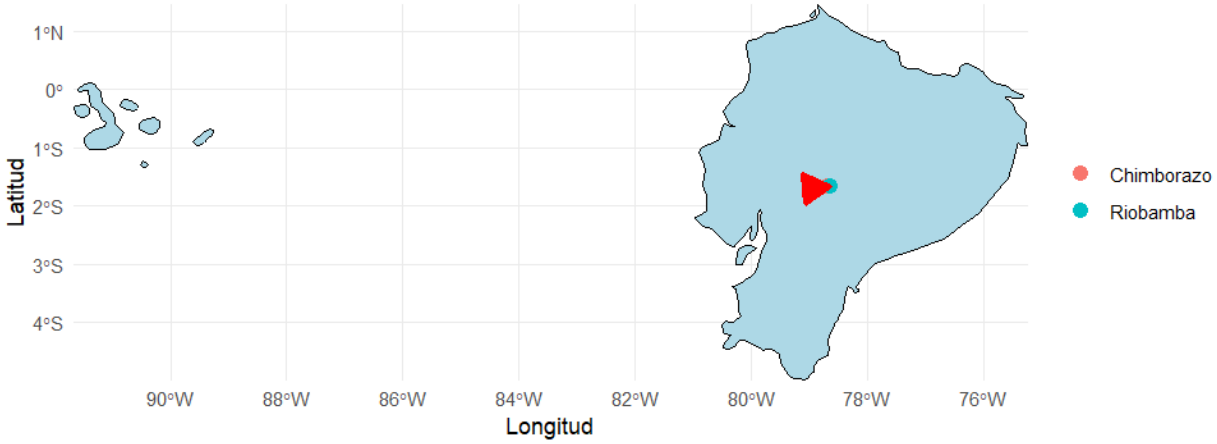
La Cooperativa de Ahorro y Crédito "Fernando Daquilema" ofrece una amplia gama de productos y servicios financieros, como cuentas de ahorro (cuenta corriente, ahorro programado), inversiones (depósito a plazo fijo) y diversos tipos de créditos (microcréditos, créditos móviles, para vivienda, consumo, agropecuarios e institucionales). Su calificación de riesgo es A- reflejando una mejora continua en la calidad de servicios e indicadores financieros y sociales, con el objetivo de brindar beneficios de calidad a sus socios y a la colectividad en general.

El Departamento de Riesgos, en conjunto con el área comercial, son responsables de gestionar y administrar la información de diversos portafolios de la cooperativa antes mencionada, siendo uno de ellos la cartera de créditos. Esta cartera contiene información tanto cualitativa como cuantitativa sobre las personas que solicitaron un préstamo, la cual constituirá la fuente de todas las variables consideradas en el presente estudio. Para esta

investigación se utilizó los datos de los socios de la agencia matriz que solicitaron un crédito durante el periodo 2023 que conforman 14774 observaciones.

Figura 5

Área de estudio.



Nota. Ubicación área de estudio COAC Fernando Daquilema Ltda. Fuente: Elaboración propia

3.2. Enfoque y tipo de investigación

Enfoque

Este estudio adopta un enfoque cuantitativo, centrado en la recopilación y análisis de datos numéricos de diversas variables mediante técnicas estadísticas (Mohajan, 2020). Este enfoque permite identificar relaciones y patrones entre variables sociodemográficas, condiciones específicas de préstamo y calificación crediticia, con el propósito de desarrollar un modelo predictivo eficiente.

La recolección y análisis objetivo de datos garantiza una evaluación rigurosa de las variables involucradas. Además, este método destaca por su precisión para identificar patrones significativos y establecer relaciones entre las variables involucrados, aspectos esenciales para construir un modelo predictivo confiable. La información obtenida proporcionará una base sólida para la toma de decisiones y la implementación de estrategias efectivas de gestión de riesgo crediticio en la institución.

Tipo de Investigación

Este estudio se caracteriza por ser una investigación correlacional, cuyo objetivo es examinar la relación entre dos o más variables y determinar cómo se interrelacionan en el contexto de la evaluación crediticia (Cauas, 2015). En este caso, se busca identificar las relaciones entre factores como el género, estado civil, condiciones de crédito, entre otras variables, y su calificación crediticia.

La investigación explicativa tiene como finalidad ofrecer una explicación y determinación de fenómenos. Dentro del contexto cuantitativo, es posible realizar estudios predictivos que permitan establecer relaciones causales entre diferentes variables y anticipar como el cambio en una variable afecta en otras (Ramos Galarza, 2020). Esta investigación aplica este método, ya que tiene como objetivo esclarecer cómo diversos factores financieros y sociodemográficas influyen en la estimación de la calificación. A través de la aplicación de estas variables comparar la eficacia de ambos modelos que optimicen la evaluación en la concesión de créditos.

3.3. Definición y operacionalización de variables

En esta investigación, la variable dependiente de la investigación se refiere a la calificación crediticia de un cliente que solicita un crédito, mientras que la variable independiente corresponde a la comparación de modelos de regresión y árboles de decisión. Las definiciones específicas se muestran en la Tabla 2.

Tabla 2

Definición de variables

Variable	Definición
Variable Independiente Comparación de modelo de regresión y árboles de decisión	Se compara principales métricas de precisión, variables significativas y facilidad de implementación en el contexto microcréditos.
Variable Dependiente: Calificación crediticia de los socios de la Cooperativa Fernando Daquilema Ltda.	Se refiere a la calificación crediticia del cliente, que indica si presenta un riesgo alto o bajo en sus obligaciones financieras.

Nota. Definición de variables de investigación. Fuente: Elaboración propia

Operacionalización de variables

Variable Independiente: Factores cualitativas y cuantitativas de los socios de la Cooperativa Fernando Daquilema Ltda. la cual se visualiza en la Tabla 3

Tabla 3

Operacionalización de variable independiente

Variable	Dimensión	Indicadores	Instrumento
Factores sociodemográficos, tipo de actividad económica, y condiciones específicas de su operación crediticia	Características personales	Edad, Sexo, Estado civil	Base de datos de la Cooperativa.
	Condiciones de Crédito	Tipo de producto, Plazo, Tasa de Interés, Estado de crédito, Colateral, Mora, Saldo capital	Base de datos de la Cooperativa
	Tipo de actividad	Empleo, Ingreso	Base de datos de la Cooperativa

Nota. Variables con relación a la estimación de la calificación crediticia según (Salazar Vergara, 2021).

La Tabla 3 presenta un conjunto de variables que podrían ser significativas para la calificación crediticia, basándose en investigaciones estudios que destacan la relevancia de estos factores en dicho proceso. Durante el desarrollo de las fases del estudio se procederá a la operacionalización de las variables específicas de la cooperativa.

La tabla en referencia se estructura en torno a una variable central desglosada en diversas dimensiones que capturan sus aspectos específicos y relevantes. Cada dimensión se evalúa mediante indicadores concretos que representan manifestaciones observables de la variable, lo que permite una medición precisa. Finalmente, se define un instrumento de recolección de datos que facilitará la obtención de información sistemática y confiable sobre los indicadores establecidos (Bauce *et al.,* 2018).

Variable Dependiente: Calificación crediticia de los socios de la Cooperativa Fernando Daquilema Ltda. la composición y su definición se encuentra en la Tabla 4.

Tabla 4

Operacionalización de la variable Dependiente

Variable	Dimensión	Indicadores	Instrumento
Calificación crediticia	Historial crediticio durante los primeros 12 meses de vida crediticia	Riesgo Alto, Riesgo Bajo	Base de datos de la Cooperativa

Nota. Definición de la variable dependiente. Fuente: Elaboración propia

3.4. Procedimientos

Para realizar esta investigación, que se enfoca en evaluar la calificación crediticia de los microcréditos concedidos por la Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda., se emplea la comparación de criterios de desempeño entre el modelo de regresión y los árboles de decisión. Por lo tanto, el proceso de investigación se divide en tres fases distintas. Cada fase se corresponde con uno de los objetivos específicos, lo que permite un análisis detallado y una evaluación profunda de los datos y de los modelos utilizados.

Fase 1. Determinación de las variables que influyen en la estimación de la calificación crediticia de los microcréditos de COAC Fernando Daquilema Ltda. en el 2023.

En la primera fase del estudio, se centra en Determinar las variables que influyen en la estimación de la calificación crediticia de los microcréditos de COAC Fernando Daquilema Ltda., comienza con la recolección de datos relacionados con el perfil sociodemográfico, actividad económica y características de los microcréditos otorgados a los socios que solicitaron un préstamo en la entidad. Esta recolección se realizará con total confidencialidad de acuerdo con las políticas internas de la institución para su relevancia en el análisis.

Una vez recolectado y para asegurar la calidad de los datos se procede a un tratamiento riguroso. Esto implica, la eliminación de valores anómalos o incompletos, conversión de variables en factores, discretización de variables. La codificación de variables categóricas se realiza en dos categorías 0 y 1 y cuando posee más de 2 categorías se utiliza la técnica de *one-hot* conocida como variables *dummy*,

En la Tabla 5 se visualiza las variables a las que se aplica estas metodologías como la recodificación, discretización y conversión a variable *dummy*. Este paso es importante, ya que asegura la precisión en el análisis y contribuye a alcanzar cada uno de los objetivos planteados. Por consiguiente, se lleva a cabo un análisis de los datos mediante técnicas estadísticas, usando tablas y gráficos para obtener una comprensión general de las características de los clientes. Esto permite identificar relaciones y patrones que puedan estar vinculados con la calificación crediticia.

Tabla 5

Recodificación de las variables

Nombre	Etiqueta	Valores	Medida	Observación
GÉNERO	Genero	1: Masculino 0: Femenino	Nominal	
ESTADO CIVIL	Estado Civil	Casado/a Soltero/a Divorciado/a Unión Libre Viudo	Nominal	Convertir a variable *Dummy*
MONTO ORIGINAL	Monto de Crédito Otorgado	Minorista (0 - 5000) Acumulación Simple (5001 - 20000) Acumulación Ampliada (20001 - 100000)	Ordinal	Convertir a variable *Dummy*
TASA DE INTERES	Tasa de Interés	1: Tasa de Interés Baja (10% - 17,50%) 0: Tasa de Interés Alta (Mayor a 18,78%)	Continua	
CONDICION OPERATIVA	Condición de Crédito	0: Legal 1: Normal	Nominal	
TIPO DE PRODUCTO	Tipo de Producto	Micro Impulso Microempresa Otro producto	Nominal	Convertir a variable *Dummy*
TOTAL DEUDA	Total Deuda	Deuda Baja (0 - 5000) Deuda Moderada (5001 - 20000) Deuda alta (Superior a 20000)	Ordinal	Convertir a variable *Dummy*
NIVEL DE RIESGO	Nivel de Riesgo	1: Alto Riesgo 0: Bajo Riesgo	Nominal	
ESTADO DE OPERACIÓN	Estado de Crédito	1: Normal 0: Novada	Nominal	
FRECUENCIA DE PAGO	Frecuencia de Pago	0: Mensual 1: Otra frecuencia	Nominal	
ACTIVIDAD ECONOMICA	Actividad Económica	Comercio Servicio Producción	Nominal	Convertir a variable *Dummy*

DESTINO FINANCIERO	Destino de Crédito	0: Activos Fijos 1: Capital de trabajo	Nominal	
ESTADO DE CREDITO	Estado de Crédito	0: Nuevo 1: Recurrente	Nominal	
TIPO DE GARANTIA	Tipo de Garantía	Sin Garante Auto Liquidables Hipotecaria	Nominal	Convertir a variable *Dummy*
CUOTAS PAGADAS	Avance de Pago de Crédito	Crédito Reciente Crédito Mediano Crédito por terminar	Nominal	Convertir a variable *Dummy*

Nota. Variables recodificadas, discretizadas y conversión a *dummy* según (Sandoval, 2015) y (Rocha Íñigo, 2020) para su posterior análisis. Fuente: Elaboración propia

Fase 2. Estimación de la calificación crediticia de los microcréditos de la COAC Fernando Daquilema Ltda. A través de la aplicación de los modelos de regresión y de árboles de decisión.

La segunda fase tiene como objetivo estimar la calificación crediticia de los microcréditos en la Cooperativa Fernando Daquilema Ltda. Esta sección inicia dividiendo el conjunto de datos en dos grupos: uno de entrenamiento, que abarca el 70% de los datos y se utiliza para ajustar los modelos, y otro de prueba, que incorpora el 30% restante para evaluar el rendimiento de los modelos desarrollados.

Se trabaja con una base de datos conformada por 14774 observaciones y 24 variables. Además, se estudia a todos los clientes donde cancelaron las 12 cuotas iniciales. Dado que, en este periodo, es posible verificar si el cliente muestra un comportamiento de pago responsable y mantiene una situación financiera estable. El modelo de regresión logística al brindar una fórmula matemática basada en la función exponencial (Ecuación 7) predice probabilidades en un rango de 0 a 1, mientras que el árbol de decisión clasifica a los clientes en categorías siguiendo la estructura de la Figura 2.

El punto de corte óptimo para la regresión logística es 0.5. Esto indica si la probabilidad es menor o igual a 0.5, el cliente se clasifica como de bajo riesgo; si es mayor a 0.5, se clasifica como de alto riesgo. Una vez obtenida las predicciones, se evalúa la precisión predictiva de cada modelo mediante matriz de confusión (Tabla 1), exactitud del modelo (Ecuación 16), tasa de error (Ecuación 17), sensibilidad (Ecuación 18), especificidad (Ecuación 19), tasa de falsos positivos (Ecuación 20), tasa de falsos negativos (Ecuación 21), pérdida logarítmica (Ecuación 22), curva ROC y área bajo la curva (AUC) (Ecuación 23).

Fase 3. Análisis de la calificación crediticia de microcréditos otorgados en la COAC Fernando Daquilema Ltda. por medio de la comparación de la precisión predictiva y facilidad de implementación entre los modelos de regresión y árboles de decisión.

La fase final de la investigación se centra en analizar la calificación crediticia de los microcréditos otorgados en la Cooperativa Fernando Daquilema Ltda. mediante la comparación de la precisión predictiva y la facilidad de implementación entre el modelo de regresión y el árbol de decisión.

Se analiza la siguiente metodología para la comparación: Preparación de los datos, la evaluación del modelo, interpretación del modelo y por último prueba del modelo que comprende la aplicación de las siguientes métricas: Matriz de confusión, exactitud del modelo, tasa de error, sensibilidad, especificidad, tasa de falsos positivos, tasa de falsos negativos, pérdida logarítmica, curva ROC, el área bajo la curva (AUC). Estas métricas proporcionan una visión clara sobre el rendimiento de cada modelo. Además, se comparan las variables significativas que influyen en la calificación crediticia.

Este análisis comparativo facilita la selección del modelo óptimo basados en su precisión, factores influyentes y facilidad de implementación y sugiere ajustes en las políticas internas según las necesidades de la cooperativa. Para llevar a cabo este estudio se utiliza software estadístico *R Studio* con librerías que permiten el desarrollo de cada etapa del análisis de datos.

CAPÍTULO IV

4. RESULTADOS Y DISCUSIÓN

En este capítulo, se presentan los resultados obtenidos en relación con cada uno de los objetivos específicos planteados. En primer lugar, se identifican y analizan las variables que influyen en la estimación de la calificación crediticia de los microcréditos otorgados por la cooperativa Fernando Daquilema Ltda. en el año 2023, evaluando su impacto y significancia dentro del modelo predictivo.

A continuación, se llevan a cabo estimaciones de las calificaciones crediticias empleando tanto el modelo de regresión y árboles de decisión, documentando minuciosamente los resultados obtenidos.

Finalmente, se realiza una comparación entre ambos modelos en términos de precisión predictiva, variables influyentes y facilidad de implementación, lo que permite determinar cuál de los dos modelos es más idóneo para estimar la calificación crediticia en microcréditos de la cooperativa.

4.1. Determinar las variables que influyen en la estimación de la calificación crediticia de los microcréditos de COAC Fernando Daquilema Ltda. en el 2023.

La base de datos de clientes que solicitaron un préstamo en la línea de microcrédito en la Cooperativa Fernando Daquilema Ltda. abarca el periodo desde enero hasta diciembre de 2023. Esta base de datos incluye información sociodemográfica y socioeconómica de los prestatarios. La información original con todas las variables y observaciones sin realizar ningún tratamiento está conformada por 14,774 registros y 45 variables.

Preparación de los datos

Bohórquez *et al.* (2020) la manipulación de los datos se centra en la eliminación de valores ausentes, la transformación y normalización de las variables, así como en el balanceo de las observaciones y la especificación del período de análisis. La base de datos completa proporcionada por la COAC Fernando Daquilema Ltda., está disponible en el siguiente archivo bdc.R. A partir de ella, se procedió a operacionalizar las diferentes variables sociodemográficas y socioeconómicas.

Datos faltantes

Los datos de la institución financiera son completos, por lo que no fue necesario emplear instrumentos adicionales para su construcción.

Transformación de variables

En la base de datos, existen variables de naturaleza cualitativa categórica. Por consiguiente, es imprescindible convertir estas variables a un formato cuantificable para su utilización en modelos de regresión y árboles de decisión. Este requisito se fundamenta en el hecho de que muchos algoritmos requieren que los datos de entrada sean numéricos para realizar predicciones de manera efectiva (Herrera Chirinos, 2024). Con el fin de realizar un análisis más efectivo de los algoritmos en este estudio, se transformarán las variables con más de tres categóricas en variables *dummy* o conocido también codificación One-hot (Rocha Íñigo, 2020). Las variables cualitativas que convirtieron a cuantitativas, siguiendo la teoría expuesta en este párrafo, se detallan en la Tabla 5.

Las variables que se eliminaron de la base de datos incluyen: Sucursal, Fecha de desembolso, Fecha de vencimiento, Código del asesor, Fecha diferido, Fecha de solicitud, Fecha de último pago, Fecha de aprobación, entre otras que no impactan en la estimación de la calificación crediticia Bohórquez *et al*. (2020). Tras eliminar estas variables, la base de datos se conforma por 24 variables, incluyendo tanto las variables originales como las transformadas a variables *dummy*, las cuales se encuentran en el archivo: bdo.R.

A continuación, se detalla la operacionalización de cada de ellas:

1. Cuotas pagadas (CUOTAS_PAGADAS): Para el análisis se filtraron todos los créditos cuyas cuotas pagadas sean superiores a los 12 meses. Luego se discretizó dividiendo el número de cuotas pagadas por el plazo total del crédito (Otra variable proporcionada por la institución) en las siguientes categorías: Crédito reciente, Crédito mediano y Crédito por terminar. Esta variable se transformó en una variable *dummy* para facilitar el análisis de los datos.

2. Monto Original (MONTOORIGINAL): Es una variable numérica definida en dólares que hace referencia al valor de la operación crediticia obtenida en la entidad financiera. Dado que esta variable tiene una amplia gama de valores, planteaba un desafío para la precisión del algoritmo, lo que a su vez restringía la

capacidad del modelo para detectar patrones más específicos en todo el rango de la variable (Sandoval, 2015). Por ende, se decidió realizar una discretización en función a las categorías establecidas por la Junta de Política y Regulación Financiera 2023, esto es: Microcrédito Minorista (0) valores entre (0 – 5000), Acumulación Simple (1) valores entre (5001 – 20000) y Acumulación ampliada (2) con valores superiores a 20000. Cabe indicar que según la Resolución No. JPRF-F-2023-086 esa categorización fue modificada en noviembre de 2023 la misma que entró en vigor desde el marzo del 2024.

3. Tasa de Interés (TASA): Esta variable de tipo numérica representa el interés o el costo en el que fue concedido un crédito. Se discretizó en tres categorías para estudiar su comportamiento estadístico y obtener soluciones exactas.

Tabla 6

Variable Tasa de Interés

Tasa de interés	
Min.	10.00
1st Qu.	17.50
Median	17.50
Mean	17.78
3rd Qu.	18.00
Max.	45.00

Nota. Estadística descriptiva de la variable tasa de Interés; Fuente: COAC Fernando Daquilema

La discretización se basó en los cuartiles obtenidos mediante análisis descriptivo de esta variable, como se visualiza la Tabla 6, y se clasificó de la siguiente manera: Todos los valores de tasa de interés entre 10% y 17.50% se clasificaron como Tasa de interés baja, los valores entre 17.50% a 17.78% como tasa de interés media y los valores superiores a 17.78 % como tasa de interés alta.

4. Total deuda (TOTALDEUDA): Variable numérica en dólares que se refiere a la deuda vigente que posee en la entidad financiera. Debido a que esta variable es de tipo cuantitativa continua posee una alta gama de valores, lo que dificulta para la

precisión del algoritmo y a la vez restringe la capacidad del modelo. Por ende, se discretizo de acuerdo con el nivel de ventas parametrizado por Banco Central del Ecuador 2015 clasificando en tres categorías: Deuda Baja, Deuda moderada y deuda alta, la cual fue convertida a la variable *dummy*.

5. Género (GÉNERO): Variable categórica que indica el tipo de género del prestatario. En este caso masculino o femenino. La codificación es (0) para el género femenino y (1) para el género masculino.

6. Estado Civil (ESTADOCIVIL): Variable que poseía 5 categorías: casado, soltero, divorciado, unión libre y viudo. Por ende, fue necesario realizar una transformación a la variable *dummy*. Estos valores toman dos valores por lo general cero (0) y (1).

7. Estado de crédito (CESTATUS): Es una variable categórica que indica el estado de un crédito, esta variable posee similitud con la variable nivel del riesgo, ya que, al calcular la correlación entre variables, se notó la presencia de alta correlación entre ambas variables. Por lo que a esta variable CESTATUS se retiró del análisis.

8. Condición Operativa (CONDICIONOPERATIVA). Es una variable categórica que indica si la operación crediticia se encuentra en estado normal o legal, la codificación es: Legal (0) y Normal (1).

9. Tipo de producto (NTIPOPRODUCTO): Es una variable categórica que indica el tipo de producto que oferta la entidad financiera las mismas que son anclados a la línea de microcréditos. En la Tabla 7***Tabla 7*** se visualiza la distribución de cada uno de los productos:

Tabla 7

Distribución de frecuencia tipo de producto

Tipo de producto	Frecuencia absoluta	Frecuencia relativa
Micro Impulso	7930	53.7
Microempresa	4280	29.0
Agropecuario	725	4.9
Daqui Esperanza	939	6.4

Conafis	17	0.1
Daqui Crecimiento	289	2.0
Randi_B2B	13	0.1
Reactivacion	340	2.3
Credi Iglesias	117	0.8
Eko Credito	109	0.7
Inclusion Emprendimiento	12	0.1
Daqui Empresario	2	0.0
Credi Listo	1	0.0

Nota. Distribución de la variable tipo de producto, para su posterior discretización. Fuente: COAC Fernando Daquilema Ltda.

Debido a la falta de suficientes observaciones en algunas categorías. Esto lleva a que el modelo de presente dificultades en las predicciones precisas (Guerra *et al.*, 2019), se decidió discretizar en tres categorías con mayor frecuencia, tal como se muestra en la Tabla 8 . Esta variable a la vez se transformó en la variable *dummy*.

Tabla 8

Distribución de frecuencias tipo de producto discretizada

Tipo de producto	Frecuencia absoluta	Frecuencia relativa
Micro Impulso	7930	53.7
Microempresa	4280	28.96
Otro producto	2564	17.35

Nota. Distribución de frecuencia de la variable tipo de producto discretizada. Fuente: COAC Fernando Daquilema.

10. Estado de operación (ESTADOPERACION): Variable categórica hace referencia al estado de la operación crediticia, se encuentra categorizado Normal (1), Novada (0).

11. Frecuencia de pago (FRECUENCIA): Variable categórica que se refiere a la forma de pago que se ajusta a la capacidad del socio. Según lo estipulado en el contrato entre el prestatario y prestamista al momento de adquisición de crédito.

Tabla 9

Distribución de frecuencia modalidad de pago

Modalidad de Pago	Frecuencia absoluta	Frecuencia relativa
Mensual	14727	99.7%
Trimestral	11	0.1%
Semanal	27	0.2%
Diaria	7	0.0%
Bi_Mensual	2	0.0%

Nota. Distribución de frecuencia de la variable frecuencia de pago para su posterior discretización. Fuente: COAC Fernando Daquilema.

Según la frecuencia de cada una de las categorías se observó que alguna de ellas aporta poca información como se muestra en la Tabla 9. Por lo que se decidió discretizar en dos categorías según Tabla 10, la codificación corresponde (0) para categoría mensual y (1) para otra frecuencia.

Tabla 10

Distribución de frecuencia modalidad de pago discretizada

Frecuencia	Frecuencia absoluta	Frecuencia relativa
Mensual	14727	99.7%
Otra_frecuencia	47	0.32%

Nota. Distribución de frecuencia de la variable frecuencia de pago discretizada. Fuente: COAC Fernando Daquilema.

12. Actividad económica (NACTIVIDAD): Esta variable refleja la actividad económica del cliente basado en su giro de negocio al momento de solicitar un crédito. Inicialmente esta variable tenía múltiples actividades la cual complicó la codificación precisa. Por tanto, se realizó una discretización en las categorías: comercio, servicio y producción. Identificando a cada actividad económica y agruparlas en las categorías mencionadas. Además, Esta variable se transformó en una variable *dummy.*

13. Destino de crédito (DESTINO): Es una variable categórica, la cual indica la actividad a la que fue destinado los prestamos adquirido. Esta variable se

categorizó según la homologación de destino de crédito del Banco Central del Ecuador 2023. Esto es, Activos Fijos (0), Capital de trabajo (1).

14. Estado de crédito (ESTADOCREDITO): Variable categórica que indica si la operación crediticia fue un crédito nuevo o un crédito recurrente, su codificación numérica es (1) recurrente y 0 para Nuevo.

15. Garantía (TIPOGARANTIA): Se trata de variable categórica que indica si la operación crediticia concedida cuenta con respaldo patrimonial, como autoliquidables, hipotecaria o a su vez sin garantía. Esta variable fue transformada en una variable *dummy.*

Periodo de estudio

Para este tipo de estudio generalmente se analiza el comportamiento del pago del cliente dentro de 3, 6, 12 meses. Según la información proporcionada por los ejecutivos de crédito, en la institución financiera como política interna establece que es necesario evaluar el comportamiento de pago en un plazo no menor a 12 meses.

Variable dependiente

El problema de clasificación surge cuando el conjunto de datos es desequilibrado, cuando el número de observaciones en un grupo es significativamente mayor que en el otro grupo. La Tabla 11 muestra la frecuencia de cada grupo, donde se observó que el riesgo normal (calificación A) tiene la mayor concentración con un 82.7%, seguido por el nivel de riesgo pérdida (Calificación E) con un 13.2% y porcentajes menores para los niveles de riesgo potencial, deficiente y dudoso.

Tabla 11

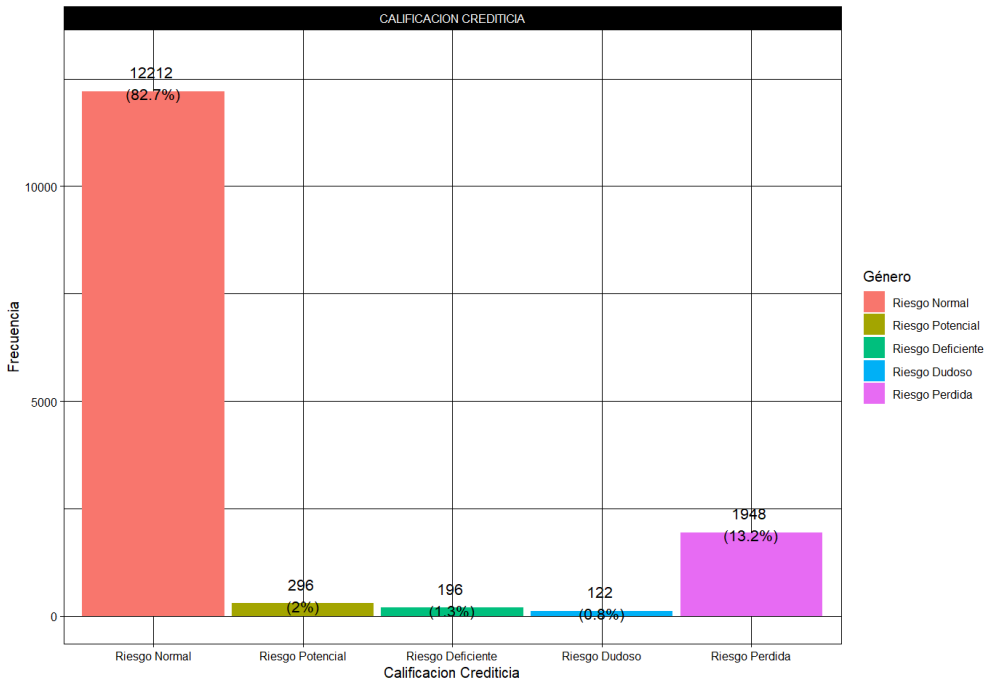
Distribución de frecuencia nivel de riesgo

Nivel de Riesgo	Total	Porcentaje
Riesgo Normal	12212	82.7%
Riesgo Potencial	296	2.0%
Riesgo Deficiente	196	1.3%
Riesgo Dudoso	122	0.8%
Riesgo Pérdida	1948	13.2%

Nota. Distribución nivel de riesgo. Fuente: COAC Fernando Daquilema.

Figura 6

Distribución de riesgo de crédito según la calificación



Nota. Distribución de nivel de riesgo según su calificación para su posterior discretización. Fuente: COAC Fernando Daquilema.

La variable dependiente se define como la probabilidad de incumplimiento, lo que implica la clasificación de los clientes en dos categorías: Bajo Riesgo (Calificación A), que incluye a aquellos clientes que cumplen con sus obligaciones a tiempo, y Alto Riesgo, que engloba a los clientes que presentan problemas debido al incumplimiento de sus obligaciones. Dentro de esta última categoría se incluyen las subcategorías de Potencial, Deficiente, Dudoso y Pérdida. La Figura 6 muestra la frecuencia de cada clase.

El objetivo de este estudio es clasificar a los clientes según su probabilidad de incumplimiento. Para ello, la variable respuesta utilizada en la predicción es la calificación crediticia, la cual se presenta en dos categorías: bajo riesgo y alto riesgo, como se ilustra en la Tabla 12 y Figura 7. La distribución de estas categorías muestra que el 82.7% de los clientes pertenece a la categoría de bajo riesgo y hace referencia a los clientes que han mantenido una buena calificación crediticia en las primeras 12 cuotas. El 17.3% restante corresponde a la categoría de alto riesgo, lo que significa que son clientes que han presentado problemas en el cumplimiento de sus obligaciones crediticias.

Tabla 12

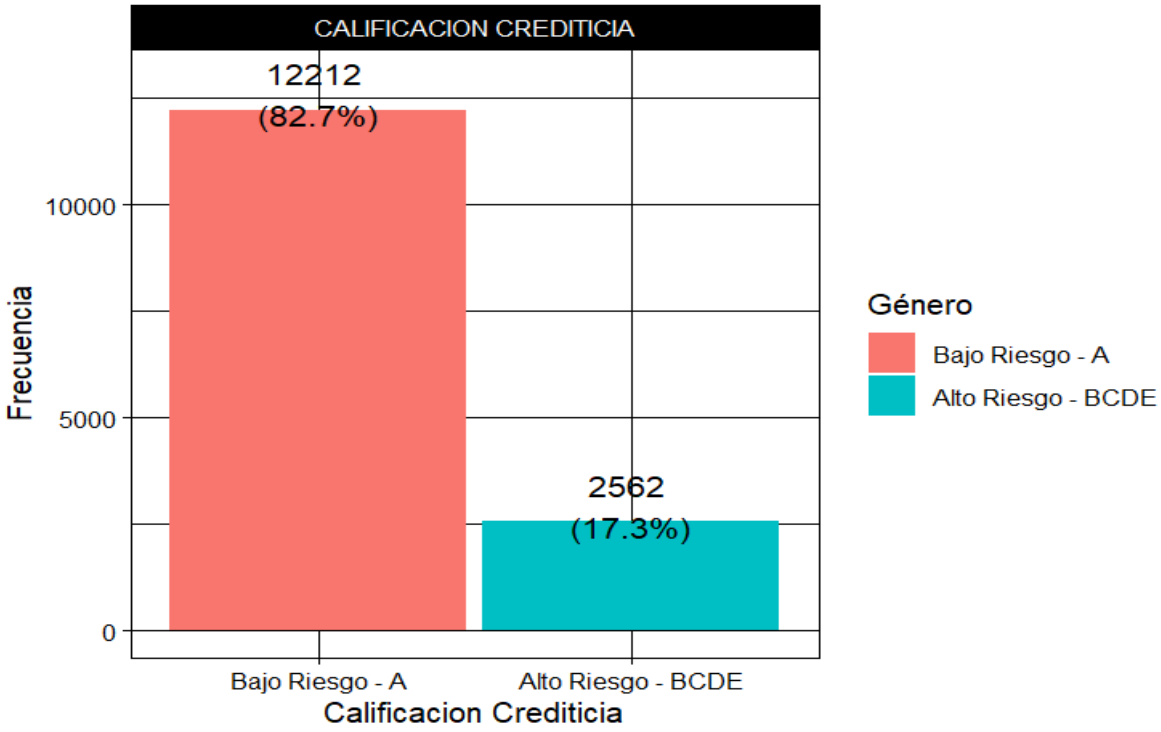
Distribución de frecuencia nivel de riesgo discretizada

Nivel de Riesgo	Total	Porcentaje
Bajo Riesgo - A	12212	82.7%
Alto Riesgo - BCDE	2562	17.3%

Nota. Distribución de frecuencia de la variable Nivel de riesgo discretizada. Fuente: COAC Fernando Daquilema.

Figura 7

Distribución de la variable dependiente



Nota. Representación de la variable dependiente discretizada. Fuente: COAC Fernando Daquilema.

En este estudio la variable calificación externa (NIVELRIESGO) se define como la variable dependiente que es la calificación crediticia, donde 1 representa alto riesgo y 0 bajo riesgo

$$Y = (y_1, y_2, \dots, y_n) \;/\; y_i = \begin{cases} \mathbf{1}, \text{si el i} - \text{ésimo cliente posee alto riesgo} \\ \mathbf{0}, \text{si el i} - \text{ésimo cliente posee bajo riesgo} \end{cases}$$

Esta variable es esencial para evaluar el desempeño de diferentes modelos que buscan determinar la probabilidad de incumplimiento de las obligaciones adquiridas por los prestatarios. La base de datos final utilizada para los modelos de regresión logística y árboles de decisión contiene 14,774 observaciones, con 23 variables independientes y una variable dependiente. Este conjunto de datos se almacenó en el archivo bdo.R.

Inicialmente, el conjunto de datos tenía 15 variables, de las cuales 14 eran independientes. Sin embargo, al transformar las variables categóricas con más de dos categorías en *dummy*, se generaron nuevas variables, resultando en un total de 23 variables independientes en la base final.

4.2. Estimar la calificación crediticia de los microcréditos de la COAC Fernando Daquilema Ltda. a través de la aplicación de los modelos de regresión y de árboles de decisión.

Partición de la base de datos:

Para ejecutar el modelo de regresión y el modelo de árboles de decisión con las variables transformadas y discretizadas, la base de datos se particionó en un conjunto de entrenamiento y un conjunto de prueba, como se muestra en la Tabla 13. Esta partición tiene el objetivo de entrenar los modelos con los datos de entrenamiento y luego evaluar su rendimiento de clasificación usando los datos de prueba.

Tabla 13

Partición de la base de datos

Datos	Alto Riesgo	Bajo Riesgo	Total	Bajo Riesgo	Alto Riesgo
Entrenamiento - 70%	1793	8548	10341	82.66%	17.34
Prueba - 30%	769	3664	4433	82.65%	17.35

Nota. Partición de la base de datos para entrenar los modelos propuestos y su posterior evaluar con datos de la prueba. Fuente: COAC Fernando Daquilema.

En el presente apartado se muestra la aplicación de los dos modelos de regresión y árboles de decisión según la información contenida en la base de datos cuyo objeto es: bdo.R.

4.2.1. Regresión Logística

1. Preparación de datos

El algoritmo de regresión logística solo acepta datos numéricos como entrada, por lo que es necesario representar las variables cualitativas de manera diferente para que el modelo pueda utilizarlas. Para las variables cualitativas con solo dos categorías, se emplea una variable binaria que toma los valores 0 o 1. Ejemplos de estas variables explicativas son: Género, tasa de interés, condición operativa, estado de operación, frecuencia de pago, destino financiero y estado de crédito. En el caso de variables cualitativas con tres o más categorías, se utilizan variables *dummy*. Algunos ejemplos de estas variables son: estado civil, monto original, tipo de producto, total deuda, actividad económica, tipo de garantía y cuotas pagadas.

2. Evaluación del modelo

En la Tabla 14 se presenta los estimadores de los coeficientes β junto con los valores de probabilidad de las variables significativas usando un nivel de significancia estadística de $\propto = 0.05$ obtenidos al entrenar el modelo de regresión logística.

Tabla 14

Variables significativas

Variable	Coeficiente estimado	Pr(>|z|)
Intercepto	3.9793602	3.56e-10
Genero1	-0.1553250	0.01723
Tasa1	-2.1602217	2e-16
Condicionoperativa1	-3.9898002	2e-16
Montooriginal_minorista1	1.9888530	2e-16
Destinofin1	0.9609872	5.48e-15
Estadocredito1	-0.4917304	3.45e-13
Montooriginal_acumulación_simple1	0.9409199	2.74e-13
Totaldeuda_moderada1	-0.6332242	6.93e-12
Cuotaspagadas_por_terminar1	-0.4731121	1.41e-09
Actividad_producción1	-0.3751895	2.71e-05
Estadocivil_soltero1	0.3566441	5.89e-07

Estadocivil_divorciado1	0.4870706	7.67e-05
Frecuencia0	-1.7755734	0.00110
Tipoproducto_microempresa1	-0.2926768	0.00434
Tipoproducto_micro_impulso1	-0.3959695	0.00579

Nota. Representación de las variables significativas en el modelo de regresión logística. Fuente: COAC Fernando Daquilema

En términos generales, de las 23 variables independientes 16 muestran coeficientes que son estadísticamente significativos debido a sus valores de p bastante bajos, lo cual indica que tienen un impacto importante sobre la variable dependiente calificación crediticia. Los coeficientes positivos de las variables indican una relación directa con la variable dependiente, mientras que los coeficientes negativos señalan una relación inversa. La baja probabilidad asociada a los valores de nivel de significancia inferiores a 0.05 refuerza la importancia de estas variables en la evaluación de la calificación crediticia, confirmando su relevancia en el modelo analizado.

Pruebas Individuales

- ✓ Planteamiento de hipótesis

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

- ✓ Nivel de significancia

$$\propto = 0.05$$

- ✓ Estadístico de prueba

$$z = \frac{x - \mu}{\sigma}$$

Ecuación 25 *Estadístico de prueba para pruebas individuales*

- ✓ Criterio de decisión

$$Si \quad p < \propto \qquad H_0 \;\; Se\ rechaza$$
$$Si \quad p > \propto \qquad H_0 \;\; No\ se\ rechaza$$

La tabla Tabla 15 contiene las 15 variables que resultaron significativas al ejecutar el modelo de regresión logística.

Tabla 15

Significancia de las variables predictoras

Variable	Estadístico_Wald	Pr(>|z|)
Intercepto	6.272192	3.56e-10
Tasa1	-13.350415	0.01723
Condición operativa1	-14.927826	2e-16
Monto original minorista1	14.171494	2e-16
Destinofin1	7.815312	2e-16
Estado credito1	-7.275416	5.48e-15
Monto original acumula simple1	7.306738	3.45e-13
Total deuda moderada1	-6.859124	2.74e-13
Cuotas pagadas por terminar1	-6.054249	6.93e-12
Actividad produccion1	-4.196555	1.41e-09
Estado civil soltero1	4.994717	2.71e-05
Estado civil divorciado1	3.954438	5.89e-07
Frecuencia0	-3.264200	7.67e-05
Tipo producto microempresa1	-2.852526	0.00110
Tipo producto micro impulso1	-2.759584	0.00434
Genero1	-2.381818	0.00579

Nota. Significancia estadística de las variables predictoras junto a su valor de probabilidad. Fuente: COAC Fernando Daquilema

En la Tabla 15, los resultados mostraron la significancia estadística de cada una de las variables para el modelo, dado que el valor de la probabilidad fue menor que el nivel de significancia $\propto = 0.05$. Por ende, se rechaza la hipótesis nula y se concluye bajo la hipótesis alternativa de que los parámetros de modelo de regresión estimada son distintos de cero, lo que sugiere que las variables consideradas en el modelo son estadísticamente significativas y aportan en la explicación de la variable dependiente.

Validación del modelo (Prueba global)

- ✓ Planteamiento de hipótesis

$$H_0: \beta_1, = \beta_2, = \beta_3, \ldots, \beta_{\mathrm{k}} \ = 0$$

$$H_1: \beta_1 \neq \beta_2 \neq \ \beta_3, \ldots, \beta_{\mathrm{k}} \ \neq 0$$

- ✓ Nivel de significancia

$$\propto = 0.05$$

- ✓ Estadístico de prueba

$$G = 2[L(\beta_i) - n_1 \ln(n_1) - n_0 \ln(n_0) + nln(n)]$$

Ecuación 26 *Estadístico G*

$$G = 3112.6$$

- ✓ Criterio de decisión

$$Si \quad p < \propto \qquad H_0 \;\; Se\ rechaza$$

$$Si \quad p > \propto \qquad H_0 \;\; No\ se\ rechaza$$

Tabla 16

Validación del modelo

N°	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	10340	9.538.891	17	3112.63	2.2e-16
2	10323	6.426.261			

Nota. Estadístico G-Desvianza junto a su valor de probabilidad. Fuente: COAC Fernando Daquilema.

En la Tabla 16, se refleja los resultados del estadístico G de razón de verosimilitud junto a su valor de probabilidad obtenida mediante la Ecuación 26, en la cual cuyo valor es menor que el valor de nivel de significancia. Por lo tanto, H_o se rechaza y no existe suficiente evidencia para decir que los coeficientes del modelo son nulos, lo que indica que al menos un β_i es diferente de cero. En consecuencia, se puede concluir que al menos unas de las variables planteadas influyen significativamente en las predicciones de la variable respuesta y se concluye que el modelo se puede estudiar.

3. Interpretación del modelo

El modelo matemático de regresión logística obtenida finalmente se aprecia en la Ecuación 27.

$$P(y_i = Riesgo\ Alto) = \frac{1}{1 + e^{-(z)}}$$

Ecuación 27 *Modelo de Regresión Logística*

Z = 3.98 – (2.16*Tasa) – (3.99*Condición Operativa) + (1.99*Monto_Original_Minorista) + (0.96*Destino_de_Crédito) – (0.49*Estado_Crédito) + (0.94*Monto_Original_Acumulación_Simple) – (0.63*Total_Deuda_Moderada) – (0.47*Cuotas_Pagadas_por_Terminar) – (0.37*Tipo_Actividad_Producción) + (0.36*Estado_Civil_Soltero) + (0.49*Estado_Civil_Divorciado) – (1.78*Frecuencia) – (0.29*Tipo_Producto_Micro_Empresa) – (0.40*Tipo_Producto_Micro_Impulso) – (0.15*Genero).

1. Tasa de interés: El valor negativo (-) del coeficiente estimado de la variable independiente TASA. Indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Es decir, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

2. Condición Operativa: El valor negativo (-) del coeficiente estimado de la variable independiente CO indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Es decir, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

3. Monto Original Minorista: El valor positivo (+) del coeficiente estimado de la variable explicativa (MOM) indica si el cliente solicita el crédito de categoría Micro minorista manteniendo constantes las demás variables, entonces aumenta la probabilidad de que el usuario pertenezca al grupo de alto riesgo.

4. Destino de crédito: El valor positivo (+) del coeficiente estimado de la variable explicativa DEST indica si el cliente solicita el crédito con destino para capital de trabajo manteniendo constantes las demás variables, entonces aumenta la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Es decir, aumenta $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

5. Estado de crédito: El valor negativo (-) del coeficiente estimado de la variable independiente ECR, indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el

cliente pertenezca al grupo de alto riesgo. Lo que indica, que disminuye $P(Y = 1)$ la probabilidad de que un cliente tenga un riesgo alto.

6. Monto original acumulación simple: El valor positivo (+) del coeficiente estimado de la variable independiente MOAS indica que, si el cliente solicita un crédito con monto que pertenece a la categoría de acumulación simple, esto aumenta la probabilidad de que el cliente pertenezca al grupo de alto riesgo. En otras palabras, aumenta $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

7. Total deuda moderada: El valor negativo (-) del coeficiente estimado de la variable independiente TDM indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Esto significa que, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

8. Cuotas pagadas por terminar: El valor negativo (-) del coeficiente estimado de la variable explicativa CPPT muestra que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. En consecuencia, disminuye $P(Y = 1)$, la probabilidad de que un cliente tenga un riesgo alto.

9. Tipo de actividad producción: El valor negativo (-) del coeficiente estimado de la variable explicativa TAPR señala que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. En otras palabras, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto

10. Estado civil soltero: El valor positivo (+) del coeficiente estimado de la variable explicativa ECS sugiere si el cliente es de estado civil soltero, manteniendo constantes las demás variables, aumenta la probabilidad de que el cliente pertenezca al grupo de alto riesgo. De este modo, aumenta $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

11. Estado civil divorciado: El valor positivo (+) del coeficiente estimado de la variable explicativa ECD, señala si el cliente es de estado civil viudo manteniendo constantes las demás variables, aumenta la probabilidad de que el cliente

pertenezca al grupo de alto riesgo. Lo que significa que, aumenta $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto

12. Frecuencia de pago: El valor negativo (-) del coeficiente estimado de la variable independiente FR indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. En otro sentido, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto

13. Tipo de producto microempresa: El valor negativo (-) del coeficiente estimado de la variable explicativa TPME indica si el cliente solicita el tipo de producto Microempresa, manteniendo constantes las demás variables, aumenta la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Esto significa que, aumenta $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

14. Tipo de producto micro impulso: El valor negativo (-) del coeficiente estimado de la variable explicativa TPMI indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. Asi que, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

15. Género: El valor negativo (-) del coeficiente estimado de la variable independiente GENERO, indica que un aumento en el promedio ponderado acumulado manteniendo constantes las demás variables, disminuirá la probabilidad de que el cliente pertenezca al grupo de alto riesgo. De este modo, disminuye $P(Y = 1)$, probabilidad de que un cliente tenga un riesgo alto.

Interpretación de los coeficientes

Tabla 17

Parámetros estimados de las variables significativas

Variable	Parámetro	Exponencial
Tasa1	-2.1602217	0.1152996
condición Operativa1	-3.9898002	0.0185034
Monto original minorista1	1.9888530	7.3071475
Destinofin1	0.9609872	2.6142761

Estado credito1	-0.4917304	0.6115672
Monto original acumulación simple1	0.9409199	2.5623373
Total deuda moderada1	-0.6332242	0.5308774
Cuotas pagadas por terminar1	-0.4731121	0.6230602
Actividad produccion1	-0.3751895	0.6871590
Estado civil soltero1	0.3566441	1.4285273
Estado civil divorciado1	0.4870706	1.6275415
Frecuencia0	-1.7755734	0.1693863
Tipo producto microempresa1	-0.2926768	0.7462633
Tipo producto micro impulso1	-0.3959695	0.6730272
Genero1	-0.1553250	0.8561369

Nota. Parámetros estimados de las variables significativas. Fuente: COAC Fernando Daquilema

En la Tabla 17 se muestran los valores de los coeficientes de las variables explicativas junto a los valores exponenciales, la interpretación de cada una de ellas se muestra a continuación:

Tasa de interés: Los clientes que tienen un préstamo con una tasa baja de interés tienen el 12% de las probabilidades (odds) de ser clasificados como de alto riesgo en comparación con los clientes que tienen un préstamo con otra tasa de interés.

Condición operativa: Los clientes que tienen un préstamo y que no presente ningún proceso judicial tienen el 2% de las probabilidades (odds) de ser clasificados como alto riesgo en comparación con los clientes que tienen procesos judiciales en las operaciones anteriores.

Monto original minorista: Los clientes que tienen un préstamo en categoría micro minorista tienen 7.3 veces más probabilidades de clasificados como de alto riesgo crediticio en comparación con los clientes que tienen un préstamo en otras categorías manteniendo todas las demás variables constantes.

Destino de crédito: Los clientes que tienen un crédito con destino para capital de trabajo tienen 2.6 veces más probabilidades de ser clasificados como de alto riesgo

crediticio en comparación con los clientes que tienen con destino para consumo, vivienda, etc. manteniendo todas las demás variables constantes.

Estado de crédito: Los clientes que son recurrentes en la institución en cuanto a las operaciones crediticias tienen el 61% de las probabilidades (odds) de ser clasificados como de alto riesgo en comparación a los clientes que son nuevos en la institución.

Monto original acumulación simple: Los clientes que tienen un préstamo en categoría de acumulación simple tienen 2.5 veces más probabilidades de ser clasificados como de alto riesgo en comparación con los clientes que tienen un préstamo en otras categorías manteniendo todas las demás variables constantes.

Total deuda moderada: Los clientes que tiene una deuda total moderada tienen el 53% de las probabilidades (odds) de ser clasificados como de alto riesgo en comparación a los clientes que poseen niveles de deuda diferentes manteniendo constantes las demás variables.

Cuotas pagadas por terminar: Los clientes que posee créditos vigentes por terminar tienen 62 % de las probabilidades (odds) de ser clasificados como de alto riesgo en comparación con los clientes que posee avance de pagos inferiores manteniendo constantes las demás variables.

Tipo de actividad producción: Los clientes que tiene una actividad económica de producción tienen el 68.71% de probabilidades (*odds*) de ser clasificados como de alto riesgo en comparación con los clientes que poseen otras actividades manteniendo todas las demás variables constantes.

Estado civil soltero: Los clientes solteros tienen 1.4 veces más probabilidades de ser clasificados como de alto riesgo crediticio en comparación con los clientes de cualquier otro estado civil manteniendo todas las demás variables constantes.

Estado civil divorciado: Los clientes divorciados tienen 1.6 veces más probabilidades de ser clasificados como de alto riesgo crediticio en comparación con los clientes de cualquier otro estado civil manteniendo todas las demás variables constantes.

Frecuencia de pago: Los clientes que posee un crédito con pago mensual tienen 16.93% de las probabilidades (*odds*) de ser clasificados como de alto riesgo en

comparación con los clientes que tienen pagos semestrales, semanales, etc. manteniendo constante todas las demás variables.

Tipo de producto microempresa: Los clientes que solicitaron un crédito de producto Microempresa tienen 74.62% de probabilidades (*odds*) de ser clasificados como de alto riesgo, en comparación con los clientes que solicitaron otros productos, manteniendo constantes todas las variables.

Tipo de producto micro impulso: Los clientes que solicitaron un credito de producto Micro Impulso tienen 67.30% de probabilidades (odds) de ser clasificados como de alto riesgo en comparación con los clientes que solicitaron otros productos manteniendo constantes todas las variables.

Género: Los clientes de género masculino tienen 85.61% de probabilidades (*odds*) de ser clasificados como de alto riesgo en comparación con los clientes de sexo femenino manteniendo constantes las demás variables.

4. Prueba del modelo

Matriz de confusión

Tabla 18

Matriz de confusión

Predicción	**Valor Real**	
	Bajo Riesgo	Alto Riesgo
Bajo Riesgo	3573	91
Alto Riesgo	454	315

Nota. Matriz de confusión regresión logística.

Fuente: COAC Fernando Daquilema

La matriz de confusión de la Tabla 18 se obtuvo utilizando un punto de corte de 0.5, un umbral comúnmente empleado. El número de clientes clasificados correctamente como alto riesgo perteneciendo realmente a dicha categoría es de 315, número de clientes clasificados como bajo riesgo, siendo realmente pertenecientes a la categoría de alto riesgo es 454, número de clientes clasificados como bajo riesgo, perteneciendo realmente a dicha categoría es de 3573, número de clientes clasificados como alto riesgo siendo realmente perteneciente a bajo riesgo es de 91

Medida de evaluación

A continuación, se muestra algunas medidas de exactitud de modelo de regresión

$$\text{error} = \frac{\text{FP} + \text{FN}}{\text{Total}} * 100 = \frac{454 + 91}{4433} * 100 = 12.3\%$$

El error mide la tasa de clasificación incorrecta. En este caso, el 12.3% de las predicciones fueron clasificados de manera incorrecta. Lo que quiere decir, el 12.3% de las veces el modelo clasificó erróneamente a los clientes en las categorías de riesgo.

$$\text{éxito} = \frac{\text{VP} + \text{VN}}{\text{Total}} * 10\frac{315 + 3573}{4433} * 100\% = 87.7\%$$

El éxito representa la tasa de clasificación correcta. En este estudio, un 87.7% de las predicciones fueron correctas. En otra palabra, el modelo acertó en la clasificación de los clientes en cuanto a riesgo crediticio.

$$\text{sensibildiad} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{315}{315 + 91} = 78\%$$

La sensibilidad mide la capacidad del modelo para identificar correctamente los casos positivos, la cual se refiere a los clientes que posee alto riesgo. El modelo identifica correctamente el 78% de los clientes que realmente posee esta categoría.

$$\text{especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{3573}{3573 + 454} = 89\%$$

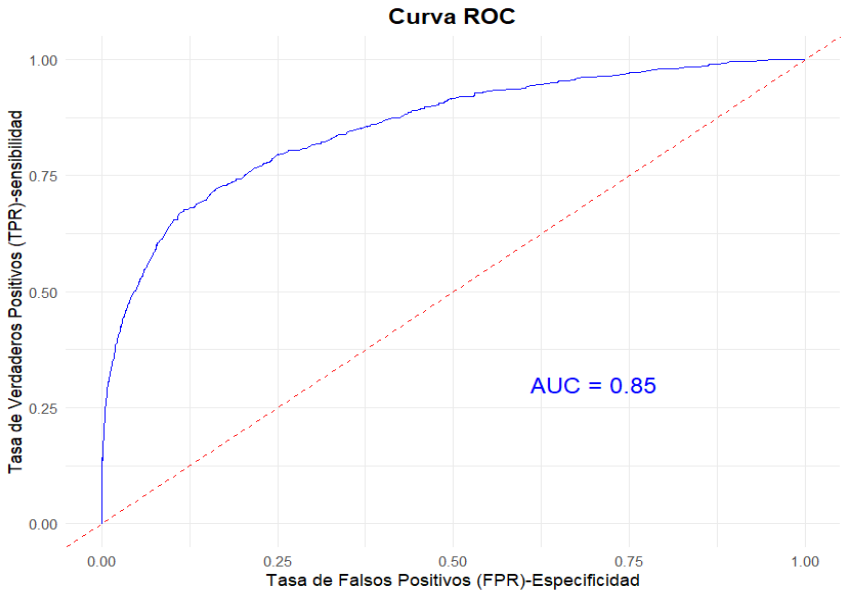
La especificidad mide la capacidad del modelo para identificar correctamente los casos negativos. En este caso, el modelo identifica correctamente el 89% de los clientes que posee bajo riesgo.

Curva Roc

A continuación, se presenta la curva ROC obtenida con los datos de prueba del modelo de regresión logística, tal como se visualiza en la Figura 8.

Figura 8

Curva ROC, datos de prueba



Nota. Curva Roc conjunto de prueba. Fuente: COAC Fernando Daquilema

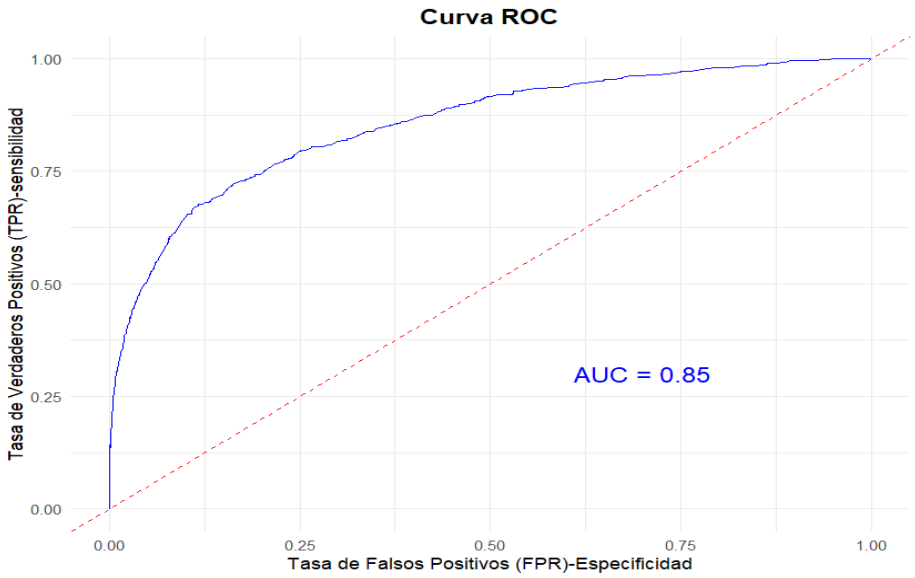
La Figura 8 de curva de ROC obtenida para el conjunto de validación en términos de tasa de falsos positivos (eje X) y las clasificaciones correctas (eje Y) se acerca a la esquina superior izquierda del gráfico, que representa una especificidad de 0.88 y una sensibilidad de 0.77. El valor de AUC para este modelo es de: 0.85 lo que indica que el modelo tiene un buen rendimiento.

Punto de corte optimo

El punto de corte en la estimación de modelos de regresión logística binaria es importante, ya que, por ese medio, se clasifica la pertenencia de un caso nuevo a una de las categorías. En la Figura 9 se muestra la curva ROC con el punto de corte 0,5.

Figura 9

Punto de corte óptimo



Nota. Curva Roc con punto de corte óptimo 0,5. Fuente: COAC Fernando Daquilema.

Una vez entrenado y verificada la significancia estadística de modelo de regresión logística, se calculó las probabilidades predichas para la base de prueba. Al inicio se estableció un punto de corte de 0,5. Según eso, los individuos con probabilidades predichas superiores a 0.5 fueron clasificadas como alto riesgo, mientras que aquellas con probabilidades inferiores se clasificaron como bajo riesgo, esto, para la construcción de la matriz de confusión y su gráfica se visualiza en la Figura 9.

4.2.2. Árboles de decisión

1. Preparación de datos

El algoritmo de árbol de decisión puede manejar datos numéricos y categóricos. Esto permite usar variables cualitativas sin transformaciones. Sin embargo, si se desea, es posible transformar los datos de entrada. Al discretizar una variable numérica busca mejorar el rendimiento del modelo, aunque no es esencial. En este estudio, se discretizaron las variables: monto original, tasa de interés, tipo de producto, total de deuda, actividad económica y cuotas pagadas.

Variable dependiente

A continuación, se muestra la Tabla 19 donde muestra la proporción de los datos en cada categoría.

Tabla 19

Distribución relativa de la variable dependiente

0: Bajo riesgo	1: Alto Riesgo
0.8266125	0.1733875

Nota. Proporción de datos. Fuente: COAC Fernando Daquilema.

La Categoría 0 se asigna a los clientes considerados de bajo riesgo, mientras que la Categoría 1 se asigna a aquellos clientes que posee alto riesgo. Es crucial señalar que la proporción de clientes en la categoría de alto riesgo es significativamente menor en comparación con bajo riesgo, como se muestra en la Tabla 19. Esta desproporción, junto con la presencia de características de alta dimensión, aumenta la complejidad del modelo. La disparidad en las proporciones de muestras positivas y negativas puede resultar en un sobreajuste, lo que disminuye la capacidad del modelo para generalizar y afecta negativamente su desempeño en la clasificación (Chen *et al.*, 2023).

En este sentido se elige un valor p proporcional para el balanceo de clase minoritaria de la variable dependiente, tal como se muestra en Tabla 20.

Tabla 20.

Valor de probabilidad

Medida	Valor de probabilidad					
	0,2	0,25	0,3	0,35	0,4	0,45
Verdadero Positivo (VP)	290	319	319	471	471	517
Falso Positivo (FP)	69	118	118	349	349	573
Falso Negativo (FN)	479	450	450	298	298	252
Verdadero Negativo (VN)	3595	3546	3546	3315	3315	3091
Sensibilidad	0,38	0,41	0,41	0,61	0,61	0,67
Especificidad	0,98	0,97	0,97	0,90	0,90	0,84
Exactitud	0,88	0,87	0,87	0,85	0,85	0,81
Error	0,12	0,13	0,13	0,15	0,15	0,19
Éxito	0,88	0,87	0,87	0,85	0,85	0,81

Log Loss	0,34	0,35	0,37	0,38	0,41	0,44
AUC	0,79	0,79	0,79	0,81	0,81	0,81

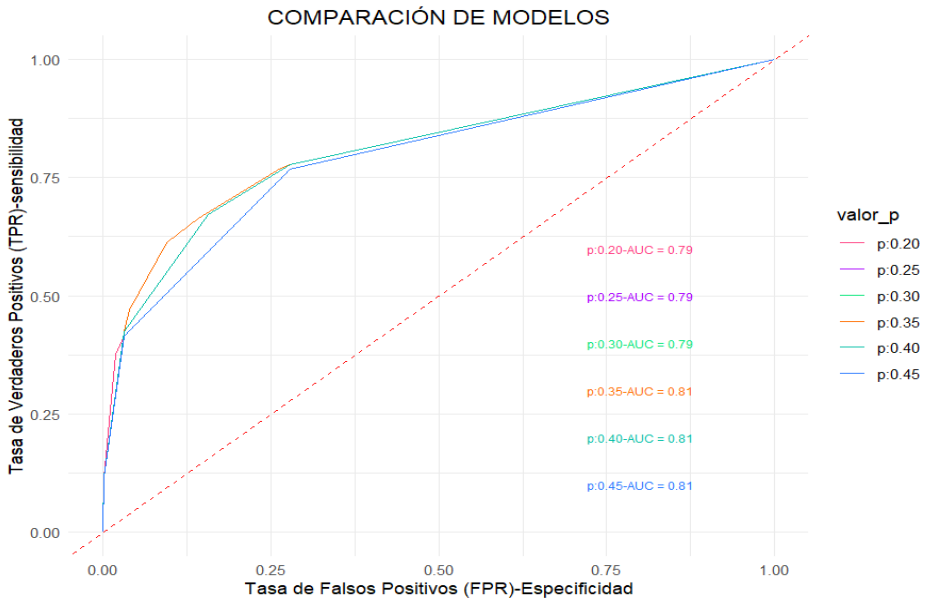
Nota. Métricas de precisión con los diferentes valores de probabilidad. Fuente: COAC Fernando Daquilema

La selección del valor de probabilidad adecuado para balancear este conjunto de dato se basa en diversas métricas de precisión. Los resultados indican que p=0.35 brinda los mejores valores en términos de exactitud, lo cual se observa en la **Tabla *20***.

A continuación, se presenta la curva ROC de diferentes valores de probabilidad, donde se visualiza que p=0.35 se encuentra por encima de los demás valores, tal como se visualiza en la Figura 10

Figura 10

Curva ROC con diferentes valores de probabilidad



Nota. Curva ROC – probabilidades. Fuente: COAC Fernando Daquilema.

Para este estudio, se emplea la técnica de *oversampling* para crear nuevas observaciones de la clase minoritaria con el objetivo de equilibrar el conjunto de datos. Los resultados se muestran en la Tabla *21*, la cual se utilizará para árboles de decisión.

Tabla 21

Frecuencia relativa de variable dependiente

0: Bajo riesgo	1: Alto Riesgo
0.5512705	0.4487295

Nota. Proporción de datos de la variable dependiente balanceado. Fuente: COAC Fernando Daquilema

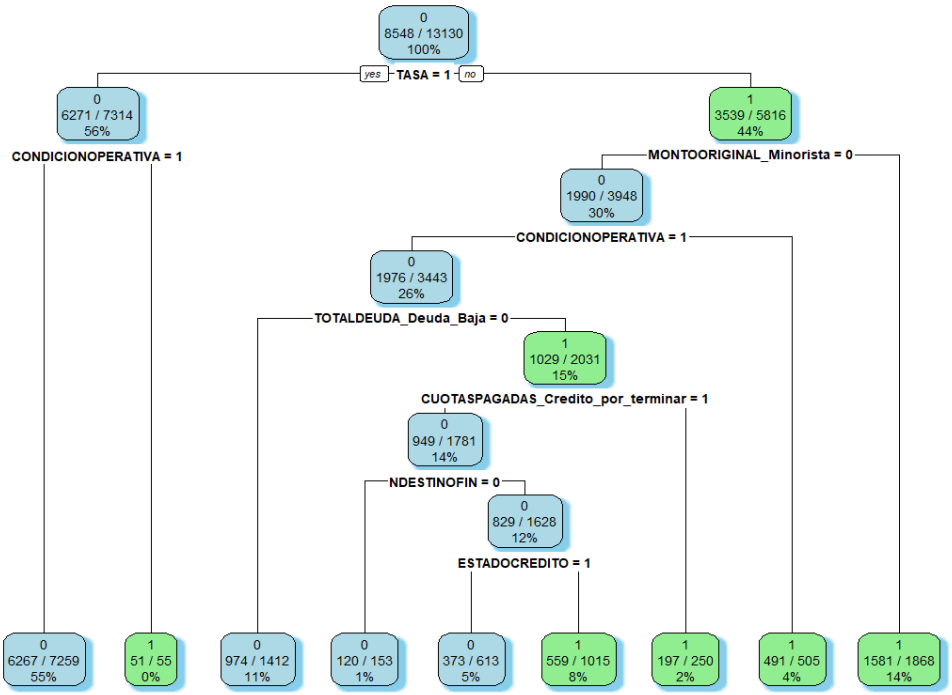
Se puede apreciar que, al establecer un umbral de probabilidad de 0.35 para la clase positiva la curva resultante supera a las curvas correspondientes a otros valores de probabilidad. Una vez balanceado los datos se obtiene el 55% de conjunto de datos para bajo riesgo y 44.87% para alto riesgo.

2. Evaluación del modelo

La explicación del árbol de decisión que se presenta en la Figura 11 es: cada nodo del árbol se muestra la categoría predicha basada en las reglas de decisión, la proporción de casos correspondientes a cada categoría, y el porcentaje de datos en dicho nodo.

Figura 11

Gráfica de modelo árbol de decisión



Nota. Árbol de decisión - datos de prueba. Fuente: COAC Fernando Daquilema

El árbol de decisión generado en la Figura 11 presenta una clasificación correcta del 85.40%. Este modelo está compuesto por 8 nodos y 9 hojas de decisión, donde se

analizan las variables significativas. Las variables con mayor capacidad de discriminación son la tasa de interés, la condición operativa, el monto original minorista y el total de deuda baja. Estas variables también demuestran su relevancia en el modelo de regresión logística, subrayando su importancia en ambos enfoques analíticos. Por ende, estas variables aportan con mayor valor de información a la variable dependiente.

Este árbol permite predecir el resultado basándose en el perfil del cliente, comenzando con variables claves como la tasa de interés, la condición operativa, y tomando en cuenta factores adicionales como el total de deuda, las cuotas pagadas y el destino de crédito. Cada nodo ofrece una probabilidad de aprobación o rechazo, lo que facilita una toma de decisiones más informada, apoyada en datos históricos de clientes similares.

3. Interpretación del modelo

El árbol de decisión obtenido, a diferencia del modelo de regresión logística, no se representa mediante una fórmula matemática, en su lugar hace uso de reglas de decisión, asociadas a las principales variables explicativas: Tasa de interés, condición operativa, monto original minorista, total deuda baja, cuotas pagadas por terminar, destino de crédito y estado de crédito. Estas reglas de decisión, las mismas que se observan en la Figura 11 del árbol de decisión, La interpretación de este modelo es la siguiente:

Nodo raíz

Todo cliente comienza desde el nodo raíz con las probabilidades iniciales de ser de la clase 0 o clase 1. La clase 0 tiene una probabilidad de aprobación 65.1%, y la clase 1 de rechazo: 34.9%

Tasa de interés

En caso de que Tasa de interés=1 que es un caso de la tasa aplicable, el árbol asume entonces la ruta con más probabilidad de aprobación del crédito. En caso de que tasa de interés =0, pasará a otra rama, la que comúnmente tiene más probabilidad de rechazo.

Condición Operativa (Tasa de interés=1)

Si la Tasa de interés es igual a 1 y la condición operativa también es igual a 1, el árbol indica una alta probabilidad de aprobación. En cambio, si la tasa de interés es igual a 1 y la condición operativa es igual a 0, la probabilidad de rechazo aumenta considerablemente.

Monto original minorista (Tasa de interés =0)

Cuando la Tasa de interés es igual a 0 y el monto del crédito solicitado se clasifica como monto original minorista igual a 0 (probablemente una cantidad menor), las probabilidades de aprobación y rechazo están más equilibradas y dependen de otras divisiones, como la condición operativa y el total de deuda baja. Si el monto original minorista es igual a 1 (posiblemente un monto mayor), la probabilidad de rechazo es alta.

Total deuda baja (Condición operativa=1)

Si total deuda baja es igual a 0, hay una alta probabilidad de aprobación, lo que sugiere que el cliente tiene poca deuda previa, facilitando así la aprobación. Si Total deuda baja es igual a 1, el árbol sigue evaluando más criterios, como las cuotas pagadas por terminar, que reflejan el estado de pagos anteriores.

Cuotas de crédito por terminar y otras variables

Cuando cuotas pagadas por terminar es igual a 1, significa que el cliente ha realizado casi todos los pagos, lo cual favorece la aprobación. Sin embargo, todavía se consideran otros factores, como el destino de crédito y el estado de crédito, antes de determinar una probabilidad final.

4. Prueba del modelo

Matriz de confusión

A continuación, se muestra la Tabla *22* que muestra la matriz de confusión resultante de modelo de árbol de decisión.

Tabla 22

Matriz confusión árboles de decisión

Predicción	Valor Real

	Bajo Riesgo	Alto Riesgo
Bajo Riesgo	3315	298
Alto Riesgo	349	471

Nota. Matriz confusión de modelo de árbol de decisión. Fuente: COAC Fernando Daquilema.

La matriz de confusión correspondiente árboles de decisión se muestra en la Tabla *22* lo que significa que 471 clientes fueron clasificados correctamente como alto riesgo perteneciendo realmente a dicha categoría, 349 clientes fueron clasificados como bajo riesgo, siendo realmente pertenecientes a la categoría de alto riesgo, 3315 clientes fueron clasificados como bajo riesgo, perteneciendo realmente a dicha categoría y 298 clientes fueron clasificados como alto riesgo siendo realmente perteneciente a bajo riesgo.

Medida de evaluación

En el siguiente apartado se muestra las medidas de exactitud de modelos de arboles de decisión

$$\mathbf{error} = \frac{\mathbf{FP} + \mathbf{FN}}{\mathbf{Total}} * \mathbf{100}\% = \frac{\mathbf{349} + \mathbf{298}}{\mathbf{4433}} * \mathbf{100}\% = \mathbf{14.60}\%$$

El 14.60% de las predicciones fueron clasificados de manera incorrecta. Lo que quiere decir, 14.60% de las veces el modelo clasificó erróneamente a los clientes en las categorías de riesgo.

$$\mathbf{éxito} = \frac{\mathbf{VP} + \mathbf{VN}}{\mathbf{Total}} * \mathbf{100} = \frac{\mathbf{471} + \mathbf{3315}}{\mathbf{4433}} * \mathbf{100}\% = \mathbf{85.40}\%$$

El éxito representa la tasa de clasificación correcta. En este estudio, el 85.40% de las predicciones fueron correctas. En otra palabra, el modelo acertó 85.40% veces en la clasificación de los clientes en cuanto a riesgo crediticio.

$$\mathbf{sensibildiad} = \frac{\mathbf{VP}}{\mathbf{VP} + \mathbf{FN}} = \frac{\mathbf{471}}{\mathbf{471} + \mathbf{298}} = \mathbf{61.25}\%$$

La sensibilidad mide la capacidad del modelo para identificar correctamente los casos positivos, la cual se refiere a los clientes que posee alto riesgo. El modelo identifica correctamente el 61.25% de los clientes que realmente posee esta categoría.

$$\mathbf{especificidad} = \frac{\mathbf{VN}}{\mathbf{VN} + \mathbf{FP}} = \frac{\mathbf{3315}}{\mathbf{3315} + \mathbf{349}} = \mathbf{90.47}\%$$

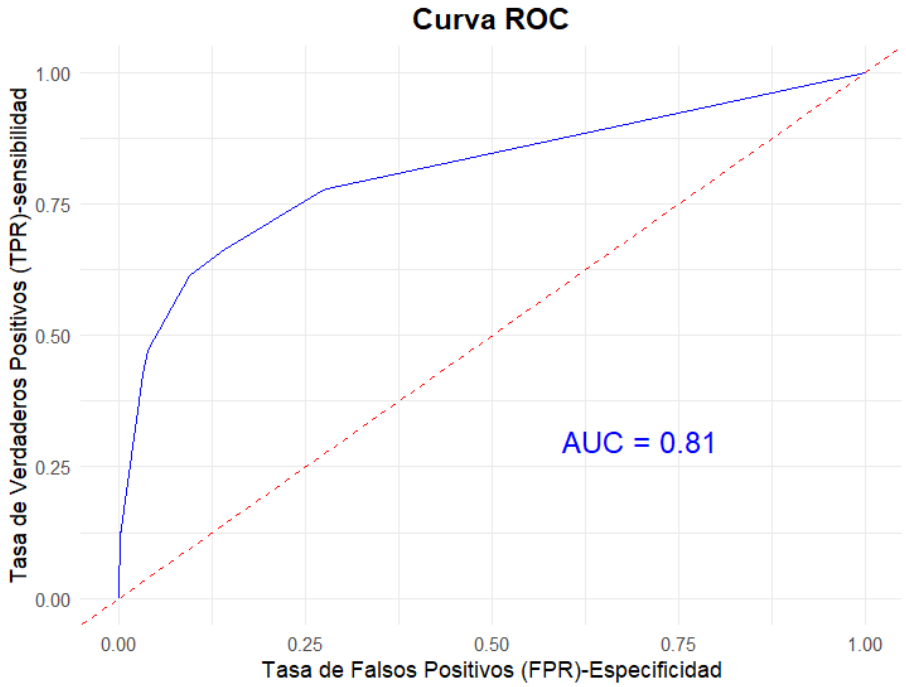
La especificidad mide la capacidad del modelo para identificar correctamente los casos negativos. En este caso, el modelo identifica correctamente el 90.47% de los clientes que posee bajo riesgo.

Curva Roc

A continuación, se presenta la curva ROC y el área bajo la curva AUC=0.81 en la Figura 12

Figura 12

Curva ROC modelo árbol de decisión



Nota. Curva ROC del modelo de árbol de decisión. Fuente: COAC Fernando Daquilema.

La gráfica de curva de ROC obtenida para el conjunto de validación en términos de tasa de falsos positivos (eje X) y la clasificación correcta (eje Y) de la matriz de confusión que se visualiza en la Tabla 22, se observa que la curva se acerca a la esquina superior izquierda del gráfico, que representa una especificidad de 0.90 y una sensibilidad de 0.61. El valor de AUC para este modelo es de: 0.81 lo que indica que el modelo tiene un buen rendimiento.

Pérdida logarítmica

La pérdida logarítmica indica la proximidad de la probabilidad de predicción con el valor real/verdadero correspondiente (0: Bajo riesgo o 1: Alto riesgo). Cuanto más diverja la probabilidad predicha del valor real, mayor será el valor de pérdida logarítmica (Dembla, 2020). Este resultado se muestra en la Tabla *23*.

En caso del presente estudio el valor de la pérdida logarítmica se obtuvo mediante la utilización del paquete *MLmetrics en R Studio,* dicho valor se describe a continuación:

Tabla 23

Pérdida Logarítmica

Métrica	Valor
Log Loss	0.3808111

Nota. Pérdida logarítmica – log los del modelo de árbol de decisión. Fuente: COAC Fernando Daquilema

La pérdida logarítmica evalúa el rendimiento de un modelo de clasificación al contrastar las probabilidades predichas con las etiquetas reales como se visualiza los resultados de matriz de confusión en la Tabla 22,. Su valor se incrementa a medida que las predicciones se alejan de las etiquetas verdaderas. Un Log Loss más bajo indica un mejor desempeño, siendo 0.00 el valor óptimo (Dembla, 2020). Para el modelo actual, la pérdida logarítmica obtenida que se muestra en la Tabla *23* es de 0.380811, lo que indica un modelo razonablemente bueno.

4.3. Analizar la calificación crediticia de los microcréditos otorgados en la COAC Fernando Daquilema Ltda. por medio de la comparación de la precisión predictiva y facilidad de implementación entre los modelos de regresión y árboles de decisión.

La selección de una metodología para la evaluación crediticia tiene que alinearse con los objetivos de cada estudio, considerando factores como alcance, tiempo y presupuesto. La decisión de un modelo sobre otro depende de diversos factores contextuales, destacando la importancia de escoger el método más apropiado para prever calificaciones de nuevos clientes. Este análisis comparativo busca ser una guía para

futuros investigadores, ayudando a evaluar diferentes modelos de clasificación y elegir el que mejor se adapte a las características de su estudio.

- ✓ **Métricas de evaluación**

En este contexto se analiza la comparación diferentes métricas de precisión para determinar el modelo de clasificación idóneo para la estimación de la calificación crediticia. Estas medias se aprecian en la Tabla 24.

Tabla 24

Métricas de Evaluación

Modelo	Exactitud	TFP I	TFN II	Auc	Log Loss	Sens.	Esp.
Regresión logística	87,71%	11,27%	22,41%	85,00%	31,00%	78%	89%
Árbol de decisión	85,40%	9,53%	38,75%	81,00%	38,00%	61%	90%

Nota. Métricas de evaluación. Fuente: COAC Fernando Daquilema.

Se evidencia que, para el conjunto de datos de prueba, el modelo de regresión logística muestra el mejor rendimiento. Esto se debe a su mayor precisión predictiva y menor porcentaje de error, alcanzando una exactitud del 87.71% y un error del 11.27%. En comparación, el modelo de árbol de decisión obtiene una exactitud del 85.40% y un error del 9.53%.

Aunque la diferencia entre ambos modelos es pequeña, se puede decir que ambos modelos proporcionan un buen rendimiento. Además, el área bajo la curva para modelo de regresión logística es de 85% frente a 81% de árboles de decisión, la pérdida logarítmica por su parte muestra el 31% para regresión logística y 38% para árboles de decisión lo cual significa que mientras más se acerca a cero el modelo es mejor. En este caso el modelo de regresión presenta una cantidad menor frente a árboles de decisión. como lo demuestra en la Tabla 24.

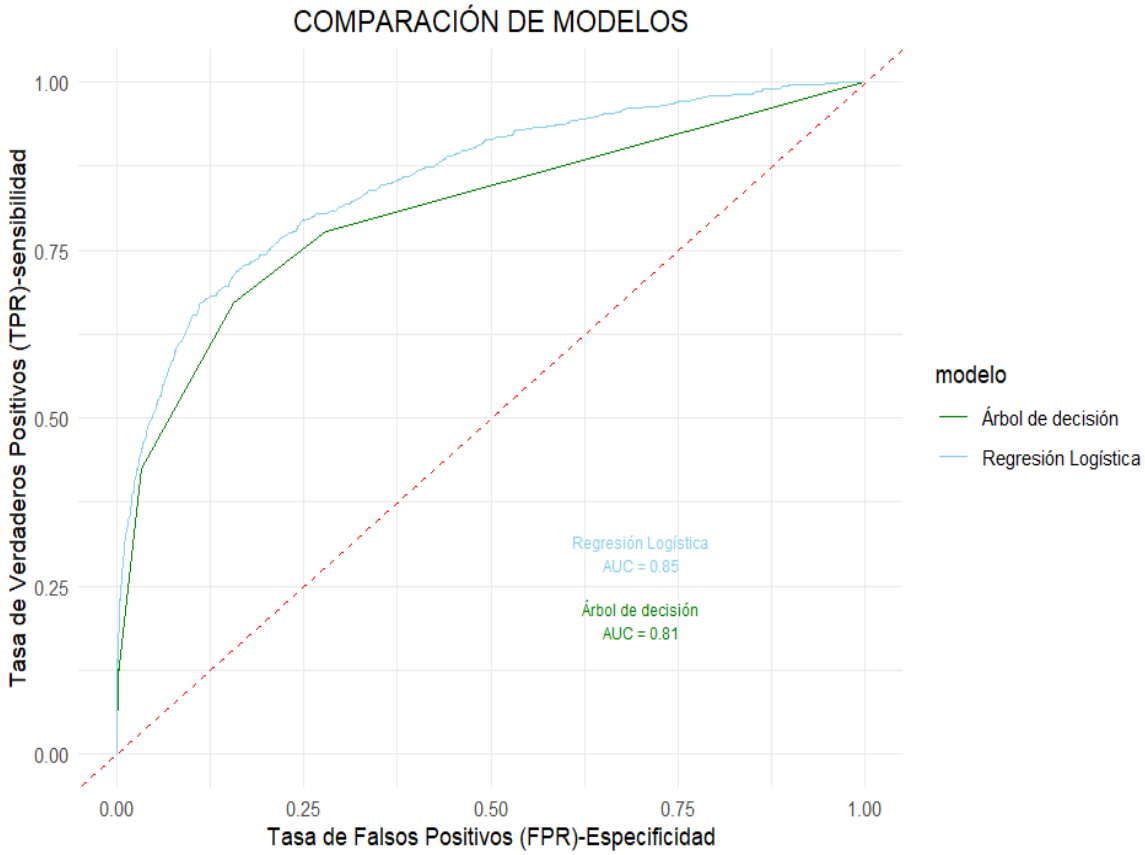
- ✓ **Curva ROC de Comparación de modelos**

En este apartado se muestra la gráfica obtenida al comparar el modelo de regresión y árboles de decisión como se muestra en la

Figura 13.

Figura 13

Comparación de modelos



Nota. Curva ROC regresión logística y árbol de decisión. Fuente: COAC Fernando Daquilema

Las curvas ROC de los modelos de regresión logística y árboles de decisión aplicados al conjunto de datos de prueba se visualiza en Figura 13. Estas curvas facilitan una comparación detallada del rendimiento de los algoritmos, evaluando la tasa de falsos positivos (eje X) y la tasa de verdaderos positivos (eje Y). Se aprecia que la curva correspondiente al modelo de regresión logística se sitúa por encima de la del modelo de árboles de decisión, lo cual sugiere un mejor desempeño en la discriminación entre las clases de los clientes que solicitan un crédito como alto riesgo y bajo riesgo.

- ✓ **Variables significativas**

Las variables que son significativas tanto en modelo de regresión y árboles de decisión se presenta en la Tabla 25.

Tabla 25

Variables Significativas

Modelo de Regresión Logística	Árboles de decisión
Género	Tasa
Tasa de interés	Condición operativa
Condición operativa	Monto original minorista
Monto original minorista	Total deuda baja
Destino de crédito	Cuotas pagadas crédito por terminar
Estado crédito	Destinofin
Monto original acumulación simple	Estado credito
Total deuda moderada1	
Cuotas pagadas crédito por terminar	
Actividad producción	
Estado civil soltero	
Estado civil divorciado	
Frecuencia	
Tipo producto microempresa	
Tipo producto micro impulso	

Nota. Comparación de variables significativas en el modelo de regresión y árbol de decisión. Fuente: COAC Fernando Daquilema.

En la Tabla 25 se presentan las variables significativas en los modelos de regresión y árbol de decisión para la estimación de la calificación crediticia de que un cliente pueda exponer un alto riesgo o bajo riesgo en acceder a un producto financiero como crédito. Las variables que son relevantes en ambos modelos son:

- Tasa de interés
- Condición operativa
- Monto original de la categoría micro minorista
- Cuotas pagadas por terminar
- Destino de crédito > Estado de crédito

Adicionalmente, se destaca que la variable total deuda resultó ser significativa únicamente en el modelo de árbol de decisión. Esto podría indicar que los árboles de decisión son más sensibles a esta variable específica en comparación con los modelos de regresión logística. Este hallazgo sugiere que, aunque ambos modelos comparten varias variables clave, hay diferencias notables en cómo cada uno de ellos puede captar y utilizar la información de ciertas variables para estimar la calificación crediticia.

Este análisis es crucial para los responsables de la toma de decisiones en las instituciones financieras, ya que proporciona una guía sobre qué variables considerar y qué modelo podría ser más adecuado dependiendo de las características de los datos disponibles.

Facilidad de implementación

Al comparar la implementación de los modelos de regresión logística y árboles de decisión, se observa que ambos tienen características distintivas que influyen en su aplicabilidad práctica. La regresión logística destaca por su simplicidad matemática y eficiencia computacional, lo que la hace adecuada para diversos entornos. Requiere menos parámetros y es menos propensa al sobreajuste con conjuntos de datos grandes y bien definidos; sin embargo, su interpretación puede ser menos intuitiva en problemas no lineales.

Por otra parte, los árboles de decisión son fáciles de implementar y manejan bien datos no lineales y variables categóricas sin mucho preprocesamiento. Su estructura en forma de árbol facilita la interpretación de los resultados, una ventaja clave para comunicar hallazgos no técnicos. No obstante, los árboles de decisión son más vulnerables al sobreajuste, especialmente con datos ruidosos o árboles profundos.

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

La correcta preparación de los datos de la cartera de microcréditos proporcionados por la cooperativa de ahorro y crédito es esencial para la aplicación efectiva de los modelos de regresión logística y árboles de decisión. El procesamiento de los datos, que abarca la limpieza, normalización, codificación y discretización de las variables, facilita la implementación de estos modelos y garantiza resultados más precisos y confiables para la aplicación en la vida real en las entidades financieras.

El uso de modelos de regresión y árboles de decisión ha facilitado la estimación de la calificación crediticia de un cliente que solicita un préstamo en la línea de microcréditos de la COAC Fernando Daquilema Ltda., determinando si representa un alto o bajo riesgo. Estos modelos han ofrecido información sobre la capacidad de los prestatarios para cumplir con sus obligaciones crediticias, aspecto fundamental para la gestión de riesgos crediticios y la toma de decisiones en la entidad financiera.

La comparación entre los modelos de regresión logística y árboles de decisión mostró diferencias clave en términos de precisión predictiva, variables influyentes y facilidad de implementación. Aunque ambos modelos lograron alta exactitud, la regresión logística destacó con un 87.71% frente al 85.40% de los árboles de decisión.

En el área bajo la curva (AUC), la regresión logística obtuvo un 85%, superando al 81% del árbol de decisión. Además, la pérdida logarítmica favoreció a la regresión logística 31% frente a 38%, indicando un mejor desempeño.

En sensibilidad, la regresión identificó correctamente al 78% de los clientes de alto riesgo, mientras que el árbol de decisión alcanzó el 61.25%. Sin embargo, en especificidad, ambos modelos fueron similares, con un 89% y 90%, destacando su capacidad para identificar a clientes de bajo riesgo.

En relación con la facilidad de implementación, la regresión logística demostró ser más eficiente en la práctica, ya que requiere menos ajustes algorítmicos para encontrar un modelo adecuado. Al contrario, los árboles de decisión necesitaron ajustes adicionales debido al desequilibrio en los datos de entrenamiento. En los factores influyentes ambos modelos señalaron variables similares que explican la estimación de la calificación

crediticia. La principal diferencia radica en que el modelo de árboles de decisión incluyó la variable explicativa Total deuda baja, la cual no resultó significativa en el modelo de regresión logística.

Con base en estas métricas, se concluye que la regresión logística es más adecuada para estimar la calificación crediticia en los microcréditos de la COAC Fernando Daquilema Ltda. Esto se debe a que sus indicadores superan al modelo de árbol de decisión, destacándose en términos de precisión y facilidad de implementación.

5.2. Recomendaciones

Seguir perfeccionando la calidad de los datos mediante el uso de técnicas avanzadas de preprocesamiento, tales como la limpieza, transformación, reducción de dimensionalidad y balanceo de datos. Asegurando de que los datos estén actualizados y sean representativos para maximizar la precisión de los modelos predictivos tanto con conjunto de entrenamiento como en prueba. En la estimación de la calificación crediticia, es esencial considerar la evolución y los pagos realizados durante al menos los últimos 12 meses. Esto requiere disponer de una base de datos con registros mensuales, permitiendo evaluar si un cliente mantiene su calificación a medida que pasa el tiempo.

Capacitar al personal en la utilización e interpretación de ambos modelos para maximizar su efectividad y elaborar estrategias para comunicar los resultados de manera eficiente a las partes interesadas, resaltando las fortalezas específicas de cada uno. Además, contemplar la incorporación de variables adicionales que puedan mejorar la capacidad predictiva de los modelos y evaluar cómo la inclusión de estas variables afecta la precisión y utilidad en la predicción de la calificación crediticia.

Se recomienda para las futuras investigaciones desarrollar una aplicación web utilizando Shiny, o cualquier paquete que permita implementar los algoritmos de los modelos en cuestión, con el objetivo de que las áreas de control puedan visualizar y comparar fácilmente ambos modelos. La aplicación proporcionará una interfaz automática que facilitará la comprensión de los resultados y permitirá aplicar estos modelos a datos reales en su vida cotidiana.

REFERENCIAS

Abdou, H. A., y Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review. *Intelligent Systems in Accounting, Finance and Management, 18*(2–3), 59–88. https://doi.org/10.1002/ISAF.325

Acuña Collazos, J. A., Domínguez Castaño, A. H., y Toro Ocampo, E. M. (2012). Una comparación entre métodos estadísticos clásicos y técnicas metaheurísticas en el modelamiento estadístico. *Revista Scientia et Technica, 2*(50), 67–76. https://dialnet.unirioja.es/servlet/articulo?codigo=4289174&info=resumen&idioma=SPA

Aji, N. A., y Dhini, A. (2019, julio 13–15). Credit scoring through data mining approach. *16th International Conference on Service Systems and Service Management*, Indonesia.

Amat Rodrigo, J. (2020). *Machine learning con R y caret*. Ciencia de Datos. https://cienciadedatos.net/documentos/41_machine_learning_con_r_y_caret

Arize AI. (2023, enero 1). Binary cross-entropy: Log loss explained. *Arize AI*. https://arize.com/blog-course/binary-cross-entropy-log-loss/

Ayús, A. L. T., Velásquez, R. E. A., y Ceballos, H. V. (2010). Estimación de las provisiones esperadas en una institución financiera utilizando modelos Logit y Probit. *Revista Ciencias Estratégicas,* 18(24), 259–270. https://www.redalyc.org/pdf/1513/151316944007.pdf

Banco Central del Ecuador. (1 de enero de 2020). *Junta de Política y Regulación Monetaria*. https://www.bce.fin.ec/junta-de-politica-y-regulacion-monetaria

Barreno-Vereau, E. (2012). Análisis Comparativo de modelos de clasificación en el estudio de la deserción universitaria. *Interfases, 5*(005), 45-82. https://doi.org/10.26439/interfases2012.n005.149

Bauce, G. J., Córdova, M. A., y Avila, A. V. (2018). Operacionalización de variables. *Revista Del Instituto Nacional de Higiene" Rafael Rangel",* 49(2). http://saber.ucv.ve/ojs/index.php/rev_inhrr/article/view/18686

Bennell, J. A., Crabbe, D., Thomas, S., y Gwilym, O. A. (2006). Modelling sovereign credit ratings: Neural networks versus ordered probit. *Expert Systems with Applications, 30*(3), 415–425. https://doi.org/10.1016/J.ESWA.2005.10.002

Bensic, M., Sarlija, N., y Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management, 13*(3), 133–150. https://doi.org/10.1002/ISAF.261

Berlanga, V., Rubio Hurtado, M. J., y Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE. Revista d'Innovació i Recerca en Educació, 6*(1), 65-79.

Boer, Y., Valencia, L., Setiadi, M. R., Setiawan, K. E., y Hasani, M. F. (2023). Classification of heart disease: Comparative analysis using KNN, random forest, Gaussian naive Bayes, XGBoost, SVM, decision tree, and logistic regression. *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–5. https://doi.org/10.1109/ICORIS60118.2023.10352195

Bohórquez, M., Torys, J., y Aguirre, M. P. (2020). Modelos de predicción de deserción de clientes para una administradora de fondos ecuatoriana. *Compendium: Cuadernos de Economía y Administración, 7*(1), 1–11. https://doi.org/10.46677/COMPENDIUM.V7I1.777

Borrero, T. D., y Bedoya, L. O. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías, 19*(4), 37–52. https://doi.org/10.18273/REVUIN.V19N4-2020004

Buitrón, A., Rodríguez, C., Calisto, M. B., y Bonilla, S. (2022). Machine learning in finance: An application of predictive models to determine the payment probability of a client. *Proceedings of the 2022 International Conference on Industrial Engineering and Operations Management*, 1–9. https://index.ieomsociety.org/index.cfm/article/view/ID/10996

Calixto Salazar, M. M., y Casaverde Carranza, L. F. (2011). Variables determinantes de la probabilidad de incumplimiento de un microcrédito en una entidad microfinanciera del Perú: Una aproximación bajo el modelo de regresión logística binaria. http://repositorio.up.edu.pe/handle/11354/1056

Caner, T. (2023). Comparison of machine learning and standard credit risk models' performances in credit risk scoring of Buy Now Pay Later customers. *METU Repository*. https://open.metu.edu.tr/handle/11511/104525

Castaño, H. F., y Ramírez, F. O. P. (2005). El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito. *Revista Ingenierías Universidad de Medellín, 4*(6), 55–75.

Causas, D. (2015). Definición de las variables, enfoque y tipo de investigación. *Biblioteca Electrónica de La Universidad Nacional de Colombia, 2*, 1–11.

Chen, H., Yang, C., Du, M., y Zhang, Y. (2023). Research on credit risk prediction under unbalanced dataset based on ensemble learning. *Mathematical Problems in Engineering, 2023*(1), 2927393. https://doi.org/10.1155/2023/2927393

Chiluiza, O., Guevara-Vega, C., Quiña-Mera, A., Landeta-López, P., y Montaluisa, J. (2023). Financial credit risk measurement using a binary classification model. *Communications in Computer and Information Science, 1705*, 241–254. https://doi.org/10.1007/978-3-031-32213-6_18

Ciencia de Datos. (2020, octubre 1). Árboles de decisión con Python. *Ciencia de Datos*. https://cienciadedatos.net/documentos/py07_arboles_decision_python.html

Congacha, S. (2023). *Análisis del riesgo en el otorgamiento de crédito de la Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda.* Universidad Nacional de Chimborazo. http://dspace.unach.edu.ec/handle/51000/10975

Curto Merino, Á. (2023). *Comparación de diferentes técnicas para predecir la pobreza.* Universidad de Valladolid. https://uvadoc.uva.es/handle/10324/63022

Dembla, G. (2020). Intuition behind log-loss score. *Towards Data Science.* https://towardsdatascience.com

Departamento Nacional de Planeación (DNP). (2023). *Consultoría para plantear recomendaciones sobre modelos de calificación crediticia para mipymes, con información alternativa* (pp. 13–14). Departamento Nacional de Planeación.

Domínguez Martin, M. (2021). *Regresión logística y técnicas de aprendizaje. Aplicaciones*. Repositorio Institucional de Documentos. https://zaguan.unizar.es/record/110300

Gabriel, J., y Vergara, S. (2021). Diseño de un modelo predictivo para otorgar créditos. *Semestre Económico*, *24*(57), 320–347. https://doi.org/10.22395/SEEC.V24N57A15

García, M. (2024). *Predicción de Default en RD: un enfoque de Machine Learning para la evaluación del riesgo crediticio*. Superintendencia de Bancos. https://sb.gob.do/publicaciones/publicaciones-tecnicas/prediccion-de-default-en-rd-un-enfoque-de-machine-learning-para-la-evaluacion-del-riesgo-crediticio/

Gonzáles, E. J., Jiménez Delgado, D. F., y Sanchez Jimenez, J. A. (2022). *Modelo de evaluación y viabilidad para préstamo de productos a los usuarios de una empresa de telefonía celular prepago en Colombia* [Tesis de pregrado, Fundación Universitaria Los Libertadores]. https://repository.libertadores.edu.co/items/b75f3613-b31f-4be0-a0a7-ab57906711d4

González, L. J. S. (2023). *Regresión logística v/s árboles de decisión en el riesgo crediticio: Logistic regression v/s decision trees in credit risk*. *RICT Revista de Investigación Científica, Tecnológica e Innovación*, 1(2), 32–37. https://doi.org/10.2992/RICT.V1I2.21

Guerra, W., Herrera, M., Fernández, L., Rodríguez Álvarez, N., Guerra, W., Herrera, M., Fernández, L., y Rodríguez Álvarez, N. (2019). Modelo de regresión categórica para el análisis e interpretación de la potencia estadística. *Cuban Journal of Agricultural Science*, *53*(1), 13–20. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2079-34802019000100013&lng=es&nrm=iso&tlng=es

Gutierrez Allende, D. E. (2023). *Stacking ensemble machine learning usando métodos de balanceo de datos para la predicción de enfermedades de la columna vertebral* [Tesis de maestría, Universidad César Vallejo]. https://repositorio.ucv.edu.pe/handle/20.500.12692/133719?locale-attribute=en

Hernández, P. A. C. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de Estadística*, *27*(2), 139–151.

Herrera Chirinos, S. E. (2024). Modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos. http://repositorio.unprg.edu.pe/handle/20.500.12893/12689

Herrera Chirinos, S. E. (2024). *Modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos* [Tesis de licenciatura, Universidad Nacional Pedro Ruiz Gallo]. http://repositorio.unprg.edu.pe/handle/20.500.12893/12689

InSight, A. (2008). *Mejores prácticas en estrategias de cobranza*. Acción InSight. https://gc.scalahed.com/recursos/files/r161r/w25484w/Mejores_Practicas_en_Estrategias_de_Cobr.pdf

Izquierdo Cruz, J. B. (2019). *Determinantes del default crediticio en una entidad del sector real para la línea de consumo motocicletas* [Tesis de maestría, Universidad del Valle]. https://bibliotecadigital.univalle.edu.co/entities/publication/de54f4ec-fdde-4c9a-b89f-4176d3a1f4dc

Junta de Política y Regulación Financiera. (2022). *Resolución JPRF-F-2022-042*. https://newsite.cite.com.ec/resolucion-jprf-f-2022-042/

Kaur, H., y Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science, 2*(2), 194–200. https://doi.org/10.3844/jcssp.2006.194.200

López, H. R. R. (2017). Diseño de un modelo matemático para la calificación de clientes morosos en una entidad comercial mediante las metodologías de árboles de decisión, análisis discriminante y regresión logística. *INNOVA Research Journal, 2*(7), 176–188. https://doi.org/10.33890/innova.v2.n7.2017.334

Lunardon, N., Menardi, G., y Torelli, N. (2014). ROSE: A package for binary imbalanced learning. *R Journal, 6*(1). https://journal.r-project.org/archive/2014/RJ-2014-008/index.html

Luque, A. G., y Peñaherrera, J. M. (2021). Cooperativas de ahorro y crédito en Ecuador: El desafío de ser cooperativas. *REVESCO: Revista de Estudios Cooperativos, 138*, 76–92.

Macías, S. E. A., y Loor, C. I. (2022). Efectos de la pandemia por Covid-19 en cooperativas de ahorro y crédito: Estudio de caso. *Cooperativismo y Desarrollo, 10*(2), 366–382. https://coodes.upr.edu.cu/index.php/coodes/article/view/506

Manvitha, T., y Rekha, K. S. (2023). Improved accuracy for prediction of leaf wetness using logistic regression algorithm compared with decision tree algorithm. *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, 1–5. https://doi.org/10.1109/ICONSTEM56934.2023.10142550

Markov, A., Seleznyova, Z., y Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science, 8*, 180–201. https://doi.org/10.1016/j.jfds.2022.07.002

Martínez Fernández, T. C. (2022). *Comparación de modelos machine learning aplicados al riesgo de crédito* [Tesis de maestría, Universidad de Concepción]. https://repositorio.udec.cl/server/api/core/bitstreams/671bc67e-081b-4f4e-ba94-291bc11e14ad/content

Méndez Gonzales, J. (2022). *Regresión dummy*. RPubs - 8. https://rpubs.com/jorge_mendez/961555

Menes Camejo, I., Arcos Medina, G., y Gallegos Carrillo, K. (2015). Revista Cubana de ciencias informáticas RCCi = Cuban journal of computer science. *Revista Cubana de Ciencias Informáticas*, *9*(4). http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992015000400008&lng=es&nrm=iso&tlng=es

Mohajan, H. K. (2020). Quantitative research: A successful investigation in natural and social sciences. *Journal of Economic Development, Environment and People*, *9*(4), 50–79.

Ortega Gutiérrez, J., Gil, J. M., Carlos, J., y Botero, V. (2010). El modelo de calificación crediticia Z-score: Aplicación del modelo Z-Score en la calificación crediticia de HB Fuller Colombia Limited. *MBA EAFIT*. https://www.eafit.edu.co/revistas/revistamba/Documents/modelo-calificacion-crediticia-z-score.pdf

Ossa, W., y Jaramillo, V. (2021). *Machine learning para la estimación del riesgo de crédito en una cartera de consumo*. Universidad EAFIT. http://hdl.handle.net/10784/29589

Otero Ortega, A. (2018). *Enfoques de investigación*. Recuperado de https://www.studocu.com/latam/document/universidad-fidelitas/metodologia-de-la-investigacion/enfoques-de-investigacion-alfredo-otero-ortega-2018/111535686

Ozturk, H., Namli, E., y Erdal, H. I. (2016). Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample. *Economic Modelling*, *54*, 469–478. https://doi.org/10.1016/J.ECONMOD.2016.01.012

Pacific Credit Rating [PCR]. (2023). *Cooperativa de Ahorro y Crédito Fernando Daquilema Ltda.* Pacific Credit Rating. https://informes.ratingspcr.com/Files/informes/1568/ec-daquilema_202212-ff.pdf

Peña, D. (2013). *Análisis de datos multivariantes.* McGraw-Hill España Cambridge.

Pérez, J. (2017). La regresión logística como modelo de predicción del riesgo crediticio en las organizaciones de la economía social y solidaria. Instituto de Investigaciones y Estudios Superiores de Las Ciencias Administrativas (IIESCA), 1–12.

Peréz, M. C., Calisto, M. B., Bonilla, S., y Riofrío, D. (2022). Application of machine learning algorithms for the prediction of payment by agreement in a debt collection company with the CRISP-DM methodology. In *3rd South American*

International Conference on Industrial Engineering and Operations Management (pp. 474–485). IEOM Society International.

Poma Salcedo, M. C. (2002). *Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación*. Universidad Nacional Mayor de San Marcos. https://sisbib.unmsm.edu.pe/bibvirtual/tesis/Basic/Salcedo_PC/Contenido.htm

Ramos Galarza, CA (2020). Los alcances de una investigación. *Revista de Divulgación Científica de la Universidad Tecnológica Indoamérica, 9* (3), 1–4. https://dialnet.unirioja.es/servlet/articulo?codigo=7746475&info=resumen&idioma=ESP

Reddy, BH y R, KP (2022). Clasificación de imágenes de fuego y humo mediante un algoritmo de árbol de decisiones en comparación con la regresión logística para medir la exactitud, precisión, recuperación y puntuación F. *2022 14. Conferencia internacional sobre matemáticas, ciencias actuariales, informática y estadística (MACS)*. https://doi.org/10.1109/MACS56771.2022.10022449

Rocha Íñigo, A. (2020) *Codificación de variables categóricas en aprendizaje automático* [Trabajo de fin de máster, Universidad de Sevilla]. https://biblus.us.es/bibing/proyectos/abreproy/71909/fichero/TFM-1909+ROCHA+I%C3%91IGO%2C+ADRI%C3%81N.pdf

Salazar Vergara, JG (2021). Diseño de un modelo predictivo para otorgar créditos. *Semestre Económico, 24* (57). https://doi.org/10.22395/SEEC.V

Sandoval, LCD (2015). *Estrategias basadas en el modelo de análisis predictivo árbol de decisión para la mejora del proceso de recaudo de cartera de la línea vehículo particular del Banco Davivienda SA* [Tesis de maestría, Pontificia Universidad Javeriana]. https://repository.javeriana.edu.co/handle/10554/16448

SEPS. (2023). *Calificación de activos de riesgo y constitución de provisiones, febrero de 2023* [Informe institucional]. Superintendencia de Economía Solidaria. https://www.seps.gob.ec/wp-content/uploads/Calificacion-activos-riesgo.pdf

Sezgin, Ö. (2006). *Métodos estadísticos en la calificación crediticia* [Tesis de maestría, Middle East Technical University]. https://open.metu.edu.tr/handle/11511/16076

Támara-Ayús, AL, Vargas-Ramírez, H., Cuartas, JJ y Chica-Arrieta, IE (2019). Regresión logística y redes neuronales como herramientas para realizar un modelo de scoring. *Revista Lasallista de Investigación, 16* (1), 187–200. https://doi.org/10.22507/rli.v16n1a5

TAŞ, C. (2023). *Comparison of machine learning and standard credit risk models' performances in credit risk scoring of buy now pay later customers* [Tesis de maestría, Middle East Technical University]. https://open.metu.edu.tr/handle/11511/104525

Tian, Z., Xiao, J., Feng, H., y Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. Procedia Computer Science, 174, 150–160. https://doi.org/10.1016/J.PROCS.2020.06.070

Tian, Z., Xiao, J., Feng, H., y Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science*, *174*, 150–160. https://doi.org/10.1016/J.PROCS.2020.06.070

Tripathi, D., Edla, D. R., Bablani, A., Shukla, A. K., y Reddy, B. R. (2021). Experimental analysis of machine learning methods for credit score classification. *Progress in Artificial Intelligence*, *10*(3), 217–243. https://doi.org/10.1007/s13748-021-00238-2

Universidad del CEMA. (2021). Modelos de Aprendizaje Automático Mediante Árboles de Decisión. *Estudios de Argentina*, Editorial UCEMA. https://ucema.edu.ar/publicaciones/download/documentos/778.pdf.

Vargas Sánchez, A., y Mostajo Castelú, S. (2014). Medición del riesgo crediticio mediante la aplicación de métodos basados en calificaciones internas. *Investigación & Desarrollo*, *2*(14), 5–25. http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2518-44312014000200002&lng=es&nrm=iso&tlng=es

Villalba, L. M. M., Camacho, F. J. A., y Ramirez, A. M. (2023). Intérprete de señales cerebrales con aprendizaje profundo para el control de Servomotores. TecnoCultura, 46.

Wang, H., Chen, W., y Da, F. (2022). Zhima credit score in default prediction for personal loans. *Procedia Computer Science*, *199*, 1478–1482. https://doi.org/10.1016/j.procs.2022.01.188

Wang, Y., Zhang, Y., Lu, Y., y Yu, X. (2020). A comparative assessment of credit risk model based on machine learning: A case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/J.PROCS.2020.06.069

ANEXO

Anexo A Certificado del Abstract por parte de Centro de Idioma

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI-
FOREIGN AND NATIVE LANGUAGES CENTER

Informe sobre el Abstract de Artículo Científico o Investigación.

Autor: Ángel Delfín Guaraca Daquilema
Fecha de recepción del abstract: 6 de febrero de 2025
Fecha de entrega del informe:6 de febrero de 2025

El presente informe validará la traducción del idioma español al inglés si alcanza un porcentaje de: 9 – 10 Excelente.

Si la traducción no está dentro de los parámetros de 9 – 10, el autor deberá realizar las observaciones presentadas en el ABSTRACT, para su posterior presentación y aprobación.

Observaciones:

Después de realizar la revisión del presente abstract, éste presenta una apropiada traducción sobre el tema planteado en el idioma Inglés. Según la rúbrica de evaluación de la traducción en Inglés, ésta alcanza un valor de 9,5; por lo cual se valida dicho trabajo.

Revisado por:	Aprobado por:
JESSICA PAOLA YANDUN BECERRA Firmado digitalmente por JESSICA PAOLA YANDUN BECERRA Fecha: 2025.02.06 15:40:00 -05'00'	JUAN CARLOS LOPEZ RUANO
Lcda. Jéssica Yandún Becerra **Docente del CIDEN**	MSc. Juan Carlos López **Coordinador de Centros Académicos y de Formación Complementaria**

Anexo B Modelo de árbol de decisión

```
print(modelo_arbol)
```

```
## n= 13130
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 13130 4582 0 (0.65102818 0.34897182)
##     2) TASA=1 7314 1043 0 (0.85739677 0.14260323)
##       4) CONDICIONOPERATIVA=1 7259  992 0 (0.86334206 0.13665794) *
##       5) CONDICIONOPERATIVA=0 55    4 1 (0.07272727 0.92727273) *
##     3) TASA=0 5816 2277 1 (0.39150619 0.60849381)
##       6) MONTOORIGINAL_Minorista=0 3948 1958 0 (0.50405268 0.49594732)
##        12) CONDICIONOPERATIVA=1 3443 1467 0 (0.57391809 0.42608191)
##          24) TOTALDEUDA_Deuda_Baja=0 1412  438 0 (0.68980170 0.31019830) *
##          25) TOTALDEUDA_Deuda_Baja=1 2031 1002 1 (0.49335303 0.50664697)
##            50) CUOTASPAGADAS_Credito_por_terminar=1 1781  832 0 (0.53284672 0.46715328)
##             100) NDESTINOFIN=0 153   33 0 (0.78431373 0.21568627) *
##             101) NDESTINOFIN=1 1628  799 0 (0.50921376 0.49078624)
##               202) ESTADOCREDITO=1 613  240 0 (0.60848287 0.39151713) *
##               203) ESTADOCREDITO=0 1015  456 1 (0.44926108 0.55073892) *
##            51) CUOTASPAGADAS_Credito_por_terminar=0 250   53 1 (0.21200000 0.78800000) *
##        13) CONDICIONOPERATIVA=0 505   14 1 (0.02772277 0.97227723) *
##       7) MONTOORIGINAL_Minorista=1 1868  287 1 (0.15364026 0.84635974)
```

Anexo C Resumen de modelo de árbol de decisión

```
summary(modelo_arbol)
```

```
## Call:
## rpart(formula = NIVELRIESGO ~ ., data = data.balanced3, method = "class")
##   n= 13130
##
##           CP nsplit rel error    xerror        xstd
## 1 0.27542558      0 1.0000000 1.0000000 0.011919893
## 2 0.05554343      1 0.7245744 0.7245744 0.010869663
## 3 0.01571366      3 0.6134876 0.6134876 0.010257978
## 4 0.01123963      5 0.5820602 0.5800960 0.010048584
## 5 0.01025753      7 0.5595810 0.5742034 0.010010297
## 6 0.01000000      8 0.5493234 0.5632911 0.009938302
##
## Variable importance
##                               TASA         NTIPOPRODUCTO_Microempresa
##                                 26                                 21
##       NTIPOPRODUCTO_Micro_Impulso                CONDICIONOPERATIVA
##                                 21                                  9
## CUOTASPAGADAS_Credito_por_terminar           MONTOORIGINAL_Minorista
##                                  7                                  6
##   MONTOORIGINAL_Acumulación_Simple                       ESTADOCREDITO
##                                  4                                  4
##             TOTALDEUDA_Deuda_Baja          TOTALDEUDA_Deuda_Moderada
##                                  1                                  1
##
## Node number 1: 13130 observations,    complexity param=0.2754256
##   predicted class=0  expected loss=0.3489718  P(node) =1
##     class counts:  8548  4582
##    probabilities: 0.651 0.349
##   left son=2 (7314 obs) right son=3 (5816 obs)
##   Primary splits:
##       TASA                         splits as  RL, improve=1406.4120, (0 missing)
##       NTIPOPRODUCTO_Micro_Impulso splits as  RL, improve= 998.6747, (0 missing)
##       NTIPOPRODUCTO_Microempresa  splits as  LR, improve= 824.5212, (0 missing)
##       CONDICIONOPERATIVA          splits as  RL, improve= 593.5050, (0 missing)
##       TOTALDEUDA_Deuda_Baja       splits as  LR, improve= 453.8849, (0 missing)
##   Surrogate splits:
##       NTIPOPRODUCTO_Microempresa         splits as  LR, agree=0.909, adj=0.795, (0 split)
##       NTIPOPRODUCTO_Micro_Impulso        splits as  RL, agree=0.907, adj=0.790, (0 split)
##       CUOTASPAGADAS_Credito_por_terminar splits as  LR, agree=0.656, adj=0.224, (0 split)
##       ESTADOCREDITO                      splits as  RL, agree=0.620, adj=0.142, (0 split)
```

```
##       CONDICIONOPERATIVA                splits as  RL, agree=0.603, adj=0.105, (0 split)
##
## Node number 2: 7314 observations,    complexity param=0.01025753
##   predicted class=0  expected loss=0.1426032  P(node) =0.5570449
##     class counts:  6271  1043
##    probabilities: 0.857 0.143
##   left son=4 (7259 obs) right son=5 (55 obs)
##   Primary splits:
##       CONDICIONOPERATIVA                splits as  RL, improve=68.24084, (0 missing)
##       MONTOORIGINAL_Minorista           splits as  LR, improve=55.66480, (0 missing)
##       TOTALDEUDA_Deuda_Baja             splits as  LR, improve=47.87843, (0 missing)
##       TOTALDEUDA_Deuda_Moderada         splits as  RL, improve=26.69557, (0 missing)
##       CUOTASPAGADAS_Credito_por_terminar splits as  LR, improve=24.26171, (0 missing)
##
## Node number 3: 5816 observations,    complexity param=0.05554343
##   predicted class=1  expected loss=0.3915062  P(node) =0.4429551
##     class counts:  2277  3539
##    probabilities: 0.392 0.608
##   left son=6 (3948 obs) right son=7 (1868 obs)
##   Primary splits:
##       MONTOORIGINAL_Minorista splits as  LR, improve=311.4000, (0 missing)
##       CONDICIONOPERATIVA      splits as  RL, improve=204.0334, (0 missing)
##       ESTADOCREDITO           splits as  RL, improve=199.8120, (0 missing)
##       TOTALDEUDA_Deuda_Baja   splits as  LR, improve=167.6278, (0 missing)
##       NDESTINOFIN             splits as  LR, improve=151.0126, (0 missing)
##   Surrogate splits:
##       MONTOORIGINAL_Acumulación_Simple splits as  RL, agree=0.884, adj=0.640, (0 split)
##       FRECUENCIA                       splits as  RL, agree=0.686, adj=0.024, (0 split)
##
## Node number 4: 7259 observations
##   predicted class=0  expected loss=0.1366579  P(node) =0.5528561
##     class counts:  6267   992
##    probabilities: 0.863 0.137
##
## Node number 5: 55 observations
##   predicted class=1  expected loss=0.07272727  P(node) =0.00418888
##     class counts:     4    51
##    probabilities: 0.073 0.927
##
## Node number 6: 3948 observations,    complexity param=0.05554343
##   predicted class=0  expected loss=0.4959473  P(node) =0.3006855
##     class counts:  1990  1958
##    probabilities: 0.504 0.496
##   left son=12 (3443 obs) right son=13 (505 obs)
##   Primary splits:
##       CONDICIONOPERATIVA                splits as  RL, improve=262.77090, (0 missing)
##       ESTADOCREDITO                     splits as  RL, improve= 79.26085, (0 missing)
##       NDESTINOFIN                       splits as  LR, improve= 64.79843, (0 missing)
##       CUOTASPAGADAS_Credito_por_terminar splits as  RL, improve= 62.67690, (0 missing)
##       MONTOORIGINAL_Acumulación_Simple   splits as  LR, improve= 60.33182, (0 missing)
##
## Node number 7: 1868 observations
##   predicted class=1  expected loss=0.1536403  P(node) =0.1422696
##     class counts:   287  1581
##    probabilities: 0.154 0.846
##
## Node number 12: 3443 observations,    complexity param=0.01571366
##   predicted class=0  expected loss=0.4260819  P(node) =0.2622239
##     class counts:  1976  1467
##    probabilities: 0.574 0.426
##   left son=24 (1412 obs) right son=25 (2031 obs)
##   Primary splits:
##       TOTALDEUDA_Deuda_Baja            splits as  LR, improve=64.28887, (0 missing)
##       ESTADOCREDITO                    splits as  RL, improve=55.36918, (0 missing)
##       MONTOORIGINAL_Acumulación_Simple splits as  LR, improve=54.36550, (0 missing)
##       TOTALDEUDA_Deuda_Moderada        splits as  RL, improve=52.11080, (0 missing)
##       NDESTINOFIN                      splits as  LR, improve=35.88495, (0 missing)
##   Surrogate splits:
##       TOTALDEUDA_Deuda_Moderada          splits as  RL, agree=0.958, adj=0.897, (0 split)
##       CUOTASPAGADAS_Credito_por_terminar splits as  LR, agree=0.751, adj=0.392, (0 split)
##       MONTOORIGINAL_Acumulación_Simple   splits as  LR, agree=0.723, adj=0.326, (0 split)
##       NDESTINOFIN                        splits as  LR, agree=0.612, adj=0.055, (0 split)
##       TIPOGARANTIA_Sin_Garante           splits as  LR, agree=0.597, adj=0.018, (0 split)
##
```

```
## Node number 13: 505 observations
##   predicted class=1  expected loss=0.02772277  P(node) =0.03846154
##     class counts:    14   491
##    probabilities: 0.028 0.972
##
## Node number 24: 1412 observations
##   predicted class=0  expected loss=0.3101983  P(node) =0.10754
##     class counts:   974   438
##    probabilities: 0.690 0.310
##
## Node number 25: 2031 observations,    complexity param=0.01571366
##   predicted class=1  expected loss=0.493353  P(node) =0.1546839
##     class counts:  1002  1029
##    probabilities: 0.493 0.507
##   left son=50 (1781 obs) right son=51 (250 obs)
##   Primary splits:
##       CUOTASPAGADAS_Credito_por_terminar splits as  RL, improve=45.13560, (0 missing)
##       ESTADOCREDITO                      splits as  RL, improve=26.44082, (0 missing)
##       NDESTINOFIN                        splits as  LR, improve=19.94789, (0 missing)
##       GENERO                             splits as  RL, improve=14.05762, (0 missing)
##       MONTOORIGINAL_Acumulación_Simple   splits as  LR, improve=13.28569, (0 missing)
##
## Node number 50: 1781 observations,    complexity param=0.01123963
##   predicted class=0  expected loss=0.4671533  P(node) =0.1356436
##     class counts:   949   832
##    probabilities: 0.533 0.467
##   left son=100 (153 obs) right son=101 (1628 obs)
##   Primary splits:
##       NDESTINOFIN                      splits as  LR, improve=21.168640, (0 missing)
##       ESTADOCREDITO                    splits as  RL, improve=20.169540, (0 missing)
##       GENERO                           splits as  RL, improve=14.989160, (0 missing)
##       MONTOORIGINAL_Acumulación_Simple splits as  LR, improve= 8.915862, (0 missing)
##       NACTIVIDAD_Produccion            splits as  RL, improve= 8.894207, (0 missing)
##
## Node number 51: 250 observations
##   predicted class=1  expected loss=0.212  P(node) =0.01904037
##     class counts:    53   197
##    probabilities: 0.212 0.788
##
## Node number 100: 153 observations
##   predicted class=0  expected loss=0.2156863  P(node) =0.0116527
##     class counts:   120    33
##    probabilities: 0.784 0.216
##
## Node number 101: 1628 observations,    complexity param=0.01123963
##   predicted class=0  expected loss=0.4907862  P(node) =0.1239909
##     class counts:   829   799
##    probabilities: 0.509 0.491
##   left son=202 (613 obs) right son=203 (1015 obs)
##   Primary splits:
##       ESTADOCREDITO                    splits as  RL, improve=19.377920, (0 missing)
##       GENERO                           splits as  RL, improve= 9.817644, (0 missing)
##       NACTIVIDAD_Produccion            splits as  RL, improve= 9.803855, (0 missing)
##       ESTADOCIVIL_Union_libre          splits as  LR, improve= 6.934211, (0 missing)
##       MONTOORIGINAL_Acumulación_Simple splits as  LR, improve= 6.724859, (0 missing)
##   Surrogate splits:
##       TIPOGARANTIA_Hipotecaria splits as  RL, agree=0.624, adj=0.002, (0 split)
##       TIPOGARANTIA_Sin_Garante splits as  LR, agree=0.624, adj=0.002, (0 split)
##
## Node number 202: 613 observations
##   predicted class=0  expected loss=0.3915171  P(node) =0.04668698
##     class counts:   373   240
##    probabilities: 0.608 0.392
##
## Node number 203: 1015 observations
##   predicted class=1  expected loss=0.4492611  P(node) =0.07730388
##     class counts:   456   559
##    probabilities: 0.449 0.551
```