Revista del Centro de Investigación de la Universidad La Salle
Vol. 16, No. 63, Enero-Junio, 2025: 4130
DOI: http://doi.org/10.26457/recein.2025.4130

Aprendizaje automático para la evaluación del riesgo crediticio en una Cooperativa de Ahorro y Préstamo

Machine learning for credit risk assessment in a Cooperative Savings and Loan

Erwis Melchor Pérez
Universidad Tecnológica de la Mixteca (México)

Moisés Emmanuel Ramírez Guzmán
Universidad Tecnológica de la Mixteca (México)

Araceli Hernández Jiménez*
Universidad del Istmo (México)

Recibido: 31 de junio de 2024
Aceptado: 06 de enero de 2025
Publicado: 14 de febrero de 2025

Resumen

El objetivo de la investigación es proponer un modelo de aprendizaje computacional que permita identificar el riesgo crediticio para las solicitudes de crédito de una Cooperativa de Ahorro y Préstamo en el Estado de Oaxaca. Para ello, se emplean los conjuntos de datos *Statlog German Credit Data* (GCD) donde el modelo de redes neuronales obtiene el mejor rendimiento con una

*Email: araceli_hdezj@hotmail.com

exactitud de 0.9533. A partir de la información proporcionada por la microfinanciera se construye un conjunto de datos, alcanzando una exactitud de 0.9243 al utilizar el clasificador *XGBoost*. Ambos conjuntos de datos se encuentran desbalanceados, por lo cual se utiliza el método de *SMOTE* para realizar sobremuestreo. El modelo propuesto permitirá reducir costos y tiempos en el análisis de solicitudes de crédito, actuando como una herramienta auxiliar para gestionar el riesgo crediticio. Esto minimizará los índices de morosidad de la microfinanciera y promoverá la inclusión financiera. A nivel regional, facilitará la identificación de patrones de comportamiento crediticio específicos, lo que permitirá ajustar las estrategias de la microfinanciera para impulsar el crecimiento económico en el estado de Oaxaca. A nivel nacional, contribuirá a la innovación y al desarrollo tecnológico en el sector financiero, fortaleciendo así el sistema financiero y, por ende, la estabilidad económica del país.

Palabras clave: Microfinanciera; Clasificadores; Red Neuronal Profunda; Preprocesamiento.

ISSN 2992-6823

Abstract

A computational learning model to identify credit risk in loan applications for a Savings and Loan Cooperative in the State of Oaxaca is proposed in this study. *Statlog German Credit Data* (GCD) datasets were used to identify credit risk for credit applications because this neural network model obtained the best performance with an accuracy of 0.9533. GCD characteristics were used as a reference to construct a dataset from the information provided by the microfinance institution. The learning model reached an accuracy rate of 0.9243, which was achieved using the *XGBoost* classifier. Both data sets were unbalanced, so the *SMOTE* method was used to oversample. The proposed model will reduce costs and time in the analysis of credit applications, acting as an auxiliary tool for managing credit risk. This will minimize microfinance delinquency rates and promote financial inclusion. At the regional level, it will facilitate the identification of specific credit behavior patterns, which will allow adjusting the microfinance institution's strategies to boost economic growth in the state of Oaxaca. At the national level, it will contribute to innovation and technological development in the financial sector, thus strengthening the financial system and, consequently, the country's economic stability.

Keywords: Microfinance institution; Classifiers; Deep neural networks; Preprocessing.

Introducción

Las entidades financieras han ido incorporando el uso de tecnología para analizar solicitudes de clientes potenciales, minimizar el riesgo crediticio, proporcionar agilidad y satisfacción en el servicio prestado. En relación al servicio de préstamos se han incorporado sistemas inteligentes para evaluar si un cliente califica o no para la aprobación de un préstamo. Rayo et al. (2010) mencionan que la banca comercial ha contado con modelos adecuados de *credit scoring* para analizar el riesgo de incumplimiento; sin embargo, no ha sido así en las Micro-Instituciones Financieras (Microfinance Institutions, MFIs, por sus siglas en inglés), señalando que la predicción del riesgo de impago debe abordarse de una manera distinta a la banca comercial, por el segmento de población al que va dirigido y pertenecer a la banca social.

En la presente investigación se estudia una sociedad cooperativa perteneciente a la familia de las SOCAP (Sociedades Cooperativas de Ahorro y Préstamo), las cuales son instituciones microfinancieras del sistema financiero mexicano, que se encuentran organizadas conforme a la Ley General de Sociedades Cooperativas y la Ley de Ahorro y Crédito, que tienen como objetivo principal contribuir a la inclusión financiera de la población de las comunidades donde no se tiene acceso a los bancos, a fin de hacerles llegar productos y servicios financieros de crédito, ahorro e inversión que contribuyan a mejorar su situación económica (CONDUSEF, 2021). Dado que el otorgamiento de créditos es una de las actividades económicas principales de las SOCAP, una consecuencia de esta actividad es el impago de los contratos de créditos (CNBV, 2017). La estimación de riesgo para cualquier crédito o préstamo monetario es realizada mediante el cálculo de la probabilidad de incumplimiento, ésta es una estimación para indicar si el contratante podrá o no cumplir con el contrato en tiempo y forma (Trejo et al., 2016).

En el contexto de la pandemia del Covid-19, la tasa de morosidad de las SOCAP se vio afectada de manera significativa (Gallardo, et al., 2023), presentando una disminución en la cartera de consumo (CNBV, 2022). Guillén & Peñafiel (2017) plantean que, la morosidad es un riesgo que cualquier institución financiera enfrenta, además del número elevado de créditos en condición de retraso o no pago, siendo una de las principales causas que atenta contra el sostenimiento de estas instituciones a largo plazo.

La importancia de la identificación del riesgo crediticio ha aumentado considerablemente con el paso del tiempo, permitiendo reducir las pérdidas y gastos debido al proceso de gestión de

ISSN 2992-6823

cobranza dentro del sector financiero, por lo tanto, una identificación temprana y adecuada representa un factor importante en la salud financiera de las entidades que emiten los créditos.

El *credit scoring* es un sistema de evaluación de las solicitudes de crédito que permite valorar de manera automática el riesgo asociado a cada solicitud. Rayo et al. (2010) lo define como un sistema que mide la probabilidad del incumplimiento del pago de un crédito que se otorga a una persona. Para la mayoría de las instituciones financieras, la adquisición de sistemas para realizar la predicción del riesgo crediticio de manera automatizada implica costos muy elevados y de difícil acceso. El análisis del riesgo crediticio en entidades financieras ha sido estudiado utilizando herramientas de aprendizaje computacional (ML, por sus siglas en inglés) y documentado por diversos autores (Lappas & Yannacopoulos, 2021; Liu et al., 2021; Zhang y Qiu, 2020; Kuppili et al., 2019).

En esta investigación se propone un modelo para predecir el incumplimiento de las solicitudes de crédito, y pueda ser de ayuda durante el proceso de su otorgamiento. La información para aplicar el modelo es proporcionada por la SOCAP objeto de estudio, a través de un convenio de colaboración, por medio del cual ha permitido a los autores del presente trabajo el acceso a su base de datos de créditos para crear el conjunto de datos, en éste los datos han sido anonimizados y preprocesados al aplicar diversas reglas internas de esta entidad sobre los mismos.

En la SOCAP, el procedimiento de identificar el riesgo de crédito se realiza de manera tradicional (utilizando hojas de cálculo que facilitan la manipulación de datos para permitir la aprobación de los créditos), por tal motivo y considerando los trabajos citados, se presenta la implementación y evaluación de los 5 clasificadores más usados en el área de aprendizaje computacional en problemas de riesgo crediticio. Para el problema presentado, los conjuntos de datos utilizados se encuentran desbalanceados, siendo éste un problema común en la mayoría de los conjuntos de datos analizados para la predicción del riesgo crediticio. Por lo tanto, para solucionar este problema se realiza el balanceo de los conjuntos de datos utilizando la técnica de sobremuestreo de la clase minoritaria con la finalidad que los modelos obtenidos no omitan esta clase durante la evaluación de los clasificadores y logre hacer una generalización de manera eficaz.

Para obtener una estimación de la calidad de los clasificadores utilizados se ha tomado como referencia el conjunto de datos *Statlog (German Credit Data)* (*GCD*), diseñado por Hofmann (1994), siendo éste un conjunto de datos ampliamente utilizado para el análisis de riesgo crediticio. Los resultados obtenidos utilizando el conjunto *GCD* superan de manera significativa los

reportados por otros autores en el estado del arte. Así mismo, se muestran los resultados de aplicar los mismos clasificadores sobre el conjunto de datos de la SOCAP.

En la sección 2 se presenta una revisión del estado del arte sobre trabajos relacionados a la predicción del riesgo crediticio, realizando un análisis de los detalles más importantes de cada propuesta. En la sección 3, se describen los conjuntos de datos utilizados y se hace un análisis de las características de ambos, también se revisan las métricas de rendimiento utilizadas para presentarlos en los resultados. En la sección 4, se presenta la metodología utilizada, así mismo, se muestran detalles del preprocesamiento hecho a los conjuntos de datos. En la sección 5, se presenta el análisis de los resultados y se hace un análisis comparativo de los resultados obtenidos al aplicar los clasificadores y diversos parámetros de ajuste seleccionados y se comparan con los obtenidos por otros autores en el estado del arte (ver Tabla 9). En la parte final se presentan las conclusiones, recomendaciones y consideraciones finales.

1. Revisión de la Literatura

Las SOCAP presentan la necesidad de identificar oportunamente a los clientes que puedan incumplir con los contratos de créditos, es por ello que Trejo-García, et al. (2016) proponen una mejora al modelo predictivo de incumplimiento utilizado por la regulación local que emite la Comisión Nacional Bancaria y de Valores (CNBV, 2017).

Shen et al. (2021) describen un método mejorado de *Synthetic Minority Oversampling Technique* (*SMOTE*, por sus siglas en inglés) propuesta por Chawla, et al. (2002), para equilibrar la cantidad de elementos de cada clase y propusieron un nuevo modelo de clasificación, resultado de la combinación de *Long Short Term Memory* (*LSTM*, por sus siglas en inglés) y el clasificador *AdaBoost.*

Por otra parte, Lappas & Yannacopoulos (2021) proponen la integración de métodos computacionales con el conocimiento de expertos en el área de riesgo de créditos. Utilizando la capacidad de interpretación y selección de características de los clasificadores sobre el conjunto de datos, la cual se ve reforzada por la participación de expertos durante el proceso de la clasificación de los créditos con la exploración de las características de los solicitantes.

Moscato et al. (2021) han utilizado un conjunto de datos de la plataforma que ofrece préstamos sociales (*Lending Club*) compuesto por 877,956 registros para realizar una evaluación por medio

ISSN 2992-6823

de herramientas de Inteligencia Artificial explicable (*XAI*, por sus siglas en inglés). Se aplicó el método SMOTE para equilibrar la distribución de las clases durante el entrenamiento de los modelos. La evaluación del conjunto de datos se hace con los clasificadores *Random Forest (RF*, por sus siglas en inglés), *Logistic Regression (LR*, por sus siglas en inglés*) y Perceptrón Multicapa* (*MLP*, por sus siglas en inglés), siendo el modelo de *LR* el que presenta el mejor rendimiento.

Liu et al. (2021) proponen el clasificador de Árbol de Decisión de Gradiente Aumentado Multigrano (*mg-GBDT*, por sus siglas en inglés) para calificar un crédito utilizando 6 conjuntos de bases de datos. Por otro lado, Zhang & Qiu (2020) presentan un modelo de puntuación de créditos basado en redes neuronales entrenadas mediante la Inteligencia de Enjambre (*SI*, por sus siglas en inglés), mostrando resultados en la búsqueda de hiper-parámetros de forma eficiente y encontrando el óptimo con una complejidad temporal adecuada, mejorando la capacidad de ajuste y generalización con una mayor precisión.

Tavana et al. (2018) proponen una Red Neuronal Artificial (*ANN*, por sus siglas en inglés) con un *MLP* compuesta por 3 capas, con una capa oculta, 9 nodos de entradas, 7 nodos en la capa oculta y un nodo en la salida, además de complementarse un modelo de Redes Bayesianas (RB). La *ANN* es utilizada para aproximar la tendencia general del riesgo y encontrar los dos factores más influyentes de forma no eficiente, la RB encuentra el factor más influyente y determina la probabilidad de que se produzca el riesgo de liquidez incluso en situaciones en las que no es posible medir todos los indicadores.

Millán & Caicedo (2018) muestran la aplicación y desempeño de tres modelos para la clasificación de las solicitudes de crédito, empleadas por las instituciones financieras en el cálculo del *scoring* de crédito, los modelos son: análisis discriminante, *LR* y *ANN*. Los resultados obtenidos favorecen en el desempeño obtenido por las *ANN* en comparación con *LR* y análisis discriminante, logrando una tasa de aciertos en la clasificación del 86.9%. Utilizando una base de datos conformada por 673 registros de una entidad financiera que incluyen las características de estado civil, edad, género, actividad económica, tipo de vivienda, dependientes económicos, plazo del préstamo, ingresos, línea de crédito y garantía.

Sun & Vasarhelyi (2018) desarrollan un sistema de predicción para identificar el riesgo de morosidad de las tarjetas de crédito de un banco brasileño. El análisis se hizo con la finalidad de explorar el rendimiento de una Red Neuronal Profunda (*DNN*, por sus siglas en inglés) en relación con las características personales y gastos de los clientes. Analizaron un total de 711,397 registros,

de los cuales 6,537 son tarjetas con morosidad y 794,860 sin morosidad. Se usó una *DNN* la cual tiene una precisión del 99.54 % y una *RN* con una precisión del 99.55 %.

Van Thiel & Van Raaij (2019) utilizaron los clasificadores *RF* y RN para identificar buenos y malos pagadores de un total de 133,152 clientes de hipotecas y tarjetas de crédito de 3 prestamistas europeos. Para el primer caso, ocuparon datos de una compañía de seguros perteneciente a una banca holandesa obteniendo un rendimiento del 95% con métodos de Inteligencia Artificial (*IA*) mejorando un 18.8 % al modelo tradicional, para el segundo caso, este proceso se realiza con datos de un banco hipotecario holandés, aplicando técnicas de *ML*, se obtiene un rendimiento del 97 % con una mejora del 21.3 % al modelo tradicional, para el tercer caso, se utilizan los datos de una compañía británica de tarjetas de crédito, obteniendo un rendimiento del 55% con una mejora del 28% sobre el modelo tradicional. Logrando predecir el riesgo de incumplimiento por medio de técnicas de *ML,* superando a los obtenidos por medio de la *LR* tradicional.

Ramos (2017) propone un método para generar modelos de clasificación de créditos, utilizando Algoritmos Genéticos (AG) para determinar el modelo de puntuaciones y *clustering* jerárquico aglomerativo para la segmentación de grupos de riesgo. Utilizando un total de 1459 pequeñas empresas, de las cuales 35 incumplieron con el pago de la deuda asumida, considerando los datos del estado financiero como lo son: pasivos, activos, ventas y utilidades. Trabajando con una población inicial de 100 individuos, con una tasa de cruzamiento del 0.7, tasa de mutación del 0.01 y 1000 generaciones. Obtuvo un porcentaje de predicciones de 58.9% de incumplimiento de pago. La desventaja de esta propuesta es que la aplicación permite al usuario manipular los datos de tal manera que pueda obtener una salida satisfactoria.

Barajas et al. (2019) desarrollan un software aplicando un AG para calcular el riesgo de incumplimiento de un individuo al solicitar un préstamo bancario, considerando las características propuestas por el banco; entre las variables analizadas destacan: edad, ocupación, nivel educativo, estado civil, género y tipo de vivienda. Estas variables coinciden con las analizadas por una gran cantidad de entidades financieras en México. Cabe mencionar que cada entidad realiza su análisis de riesgo crediticio con parámetros que ellos definen y en relación a la información solicitada por las entidades reguladoras.

Bussmann et al. (2020) proponen un modelo de *ML* que puede utilizarse en la gestión del riesgo crediticio y, en particular, en la medición de los riesgos que surgen cuando se solicita un crédito empleando plataformas de préstamos. Con una base de datos con 15,000 registros de pequeñas y

ISSN 2992-6823

medianas empresas que solicitan crédito, compuesta por registros que son morosos y no morosos, clasificándose de acuerdo con un conjunto de características financieras similares, que pueden emplearse para explicar su calificación crediticia para predecir su comportamiento futuro. El modelo de *RL* obtiene un 80% de precisión; sin embargo, el modelo *XGBoost* resulta tener una mejor eficiencia del 93%.

Ossa & Jaramillo (2021) muestran el desempeño de precisión de un modelo de *LR* frente a algunos modelos de aprendizaje computacional para la estimación de riesgo crediticio. Utilizando los modelos de *RL, RF, Support Vector Machine* (SVM, por sus siglas en inglés) y *MLP*, compararon la estimación de los clientes que van a caer en mora, y concluyen que el modelo más equilibrado al momento de realizar la evaluación es *RF*, el cual presenta el mejor ajuste con las métricas de exactitud evaluadas. Evaluando una base de datos compuesta por 15,060 registros con 18 variables, entre los cuales se encuentran: edad, monto del crédito, plazo, ingresos, gastos, residencia, tiempo en trabajo, etc.

Guevara & Freire (2021) proponen desarrollar un modelo de clasificación de clientes en Insotec que permita disminuir el riesgo crediticio y mejore los tiempos de respuesta a las solicitudes, evidenciando el éxito del clasificador *RF* con una precisión de efectividad de 97.2% y una tasa de error de 2.8%. Se utilizaron datos que se clasifican en morosos y no morosos, compuesto por 18 variables y 63,896 registros. Dentro de los cuales se encuentran: tipo de préstamo, año y mes de última transacción, género, estado civil, grado de estudio, edad, patrimonio, ingreso y egresos.

Tripathi et al. (2020) utilizan una Máquina de Aprendizaje Extremo (*ELM*, por sus siglas en inglés) como una herramienta de clasificación para la evaluación del riesgo de crédito. Proponiendo una nueva función de activación algebraica (ecuación 1) y un enfoque evolutivo para obtener los pesos y sesgos utilizando el algoritmo de optimización *Bat*. Para realizar la validación del modelo se utilizaron los conjuntos de datos Alemán[2] (GCD) realizado por Hofmann (1994), el conjunto de datos Australiano[3] *Australian Credit Approval* (ACA) proporcionado por Ross (1987) y el conjunto de datos Japonés[4] *Japanese Credit Screening* (*JCS*) creado por Sano (1992).

[2] https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[3] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29

[4] https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening

$$f_p(x) = \frac{x}{(1 + x^{(4p-2)})^{\frac{1}{(4p-2)}}}, p = 1,2,3,..,n \tag{1}$$

Kuppili et al. (2019) utilizan una Máquina de Aprendizaje Extremo con Picos (*SELM*, por sus siglas en inglés) con una función generadora de picos en el Modelo de Integración y Disparo No-lineal (*LNIF*, por sus siglas en inglés), validando el modelo con los conjuntos de datos *GCD*, *JCS*, *ACA* y *Bankruptcy dataset* (*BKD*), además de utilizar el método de Análisis de Componentes Principales (*PCA*, por sus siglas en inglés) para la selección de aquellas características con la mayor varianza.

Zhu et al. (2023) proponen un método híbrido *CNN-LightGBM*, el cual combina la *Convolutional Neural Network (CNN*, por sus siglas en inglés*)* y el clasificador *LightGBM* para establecer un modelo de predicción de riesgo crediticio. La *CNN* es utilizada para la extracción de características para después entrenar el modelo *LightGBM*. Así mismo, Li et al. (2023) utilizan los algoritmos de agrupamiento para identificar grupos de riesgo crediticio y analizar subpatrones, además de utilizar el método *NystromNet*, el cual consta de 2 etapas, la primera agrupa datos y la segunda es enfocada a la separabilidad en dos clases de los datos. Mientras que Wang & Zhang (2023) construyen un nuevo modelo de aprendizaje computacional utilizando *CNN* y utilizando la validación cruzada para el aumento del rendimiento del modelo.

Por otro lado, Pérez et al. (2023) se enfocan en la creación de una metodología para obtener la mejor heurística de los modelos en la función de pérdida durante el entrenamiento de modelos para predecir el riesgo crediticio. Yang et al. (2023) proponen un clasificador basado en *RN* para identificar el riesgo crediticio y seleccionar las características más significativas en conjunto de datos de alta dimensionalidad.

Como se puede observar en el estado del arte existen diversos conjuntos de datos que son utilizados para la predicción del riesgo crediticio. Cada conjunto de datos contiene diferente número de características, registros y tipo de datos. Sin embargo, realizando un análisis a cada conjunto de datos, se concluye que la base de datos *GCD*, presenta características similares a la información proporcionada por la SOCAP objeto de estudio.

En la documentación de otros conjuntos (*JCS* y *ACA*) no se realiza la descripción de cada una de sus características utilizadas para la predicción del riesgo de crédito. De esta manera y considerando los trabajos anteriormente citados, se presenta la implementación de 5 clasificadores que han sido utilizados ampliamente para evaluar las solicitudes de crédito. Una característica de

ISSN 2992-6823

estos conjuntos de datos es el desbalance en el número de elementos de cada clase, por lo tanto, para solucionar esta condición se utiliza la técnica de sobremuestreo de la clase minoritaria con la finalidad de que los clasificadores no omitan dicha clase durante la evaluación y logre hacer una generalización de manera más eficiente de los registros presentados.

Con base en la revisión del estado del arte, es posible obtener un margen de error bajo en el análisis de *score* crediticio al utilizar modelos de aprendizaje computacional. La obtención de estos modelos tiene por objetivo ser una herramienta eficaz para el análisis de solicitudes de crédito para las SOCAP.

2. Materiales y métodos

Las SOCAP presentan problemas al momento de realizar la identificación de los socios que son aptos para el otorgamiento de un crédito a través del análisis e interpretación de sus datos personales, bancarios y económicos, aplicando los lineamientos establecidos por la regulación vigente. La recopilación de información sobre los clientes es esencial en el análisis de riesgo crediticio para el sector financiero. En el resto de esta sección se describe el proceso de selección de datos, utilizando el conjunto de datos *GCD* y la construcción del conjunto de datos de la SOCAP.

2.1. Base de datos

El conjunto *GCD* ha sido utilizado para el análisis del riesgo de crédito por diversos autores en sus trabajos de investigación donde destacan: (*Dharwadkar & Pantil, 2018)* usan ANN, SVM y DNN obteniendo exactitudes de 72%, 72% y 76%, respectivamente; (Brown & Mues, 2012) usan diversas técnicas, dentro de las que destacan *RF* y Mejora de gradiente (gradient boosting) especialmente en conjuntos desbalanceados; (*Pandey et al., 2017)* usaron 9 modelos de clasificadores, de los cuales la *ELM* tuvo mayor exactitud (96.33%); así mismo, (West, 2000) reportó el uso de 5 tipos de arquitecturas de redes neuronales obteniendo resultados de hasta 78% de exactitud. Este conjunto de datos tiene dos clases, que pueden ser créditos que representan un alto o un bajo riesgo; contiene características del cliente como lo son la cuenta corriente, el historial de crédito, finalidad del crédito, monto del crédito, cuenta de ahorro, antigüedad en el empleo, estado civil, género, tiempo de residencia, propiedad, edad, número de créditos en el banco, empleo, teléfono y si es trabajador extranjero. El conjunto está compuesto por 20 características y 1,000 registros divididos en 700 de alto riesgo y 300 de bajo riesgo.

La selección eficiente de las características o atributos para realizar el análisis del riesgo crediticio dentro de las SOCAP son de suma importancia, estas características constituyen la base fundamental que implica otorgar o negar una solicitud de préstamo. Con la finalidad de realizar una predicción del riesgo crédito en las SOCAP del Estado de Oaxaca, se realizó un convenio de participación con la Cooperativa para proporcionar de manera anonimizada sus datos, durante este proceso se llevó a cabo una serie de pasos para preprocesar el conjunto de datos.

Como primer paso, se exportó el conjunto de datos inicial a una base de datos *PostgreSQL* para facilitar el tratamiento de los datos por medio de las funciones propias del gestor de base de datos. El siguiente paso fue eliminar aquellos datos atípicos o aquellos registros con información incompleta o vacíos y se eliminaron los registros duplicados (ver Figura 1); basado en la información proporcionada por la SOCAP, se identificaron múltiples registros que hacían referencia a la misma solicitud en donde la única diferencia era la garantía y su monto, en este caso se unificaron todos los registros, quedando un solo registro con la información de las garantías y la suma de sus montos.

Figura 1

Nota. Elaboración propia utilizada para el procesamiento de datos de la SOCAP.

El resultado de este proceso es la obtención del conjunto de datos limpios y que contiene la información de los socios de la entidad financiera, del período comprendido desde el 1 de enero del 2018 hasta el 31 de marzo del 2022. La determinación de la lista de características del conjunto de la SOCAP fue definida con base a la experiencia de los expertos de la institución financiera, políticas de la institución y a la información utilizada por otros autores durante la revisión del estado del arte. El conjunto de datos *GCD* presenta una descripción clara de los atributos utilizados a diferencia de otros conjuntos de datos mencionados en el estado del arte, donde existe poca o nula descripción de la información de las características utilizadas. Este proceso generó una lista de las características de mayor peso para identificar si una persona es viable para ser acreedor o no de un crédito, las cuales son: el código postal, ocupación, relación de ingresos y egresos, plazo y monto del crédito e historial crediticio (este dato proviene de la consulta obtenida del buró de

ISSN 2992-6823

crédito). Al finalizar la construcción de la base de datos, esta quedó compuesta por 23 características y 5,237 registros.

Además, con la finalidad de identificar la relación entre las variables independientes y dependientes, se realiza un análisis de correlación mostrada en el Apéndice B, donde se muestra el análisis estadístico de Chi-cuadrada y ANOVA de ambas bases de datos utilizadas para la predicción del riesgo crediticio. Así mismo, se identifica que estas variables coinciden con aquellas de mayor peso en el análisis por parte de los expertos de la institución financiera al momento de determinar si una persona es acreedora o no a un préstamo. Estos métodos ayudan a identificar aquellas características redundantes y de menor importancia al momento de realizar una predicción del riesgo crediticio y evitar el sesgo hacia una clase del conjunto de datos con los modelos de *ML*.

En la Tabla 1 se presenta la descripción de los conjuntos de datos utilizados durante el desarrollo de este trabajo. La columna Conjunto de datos presenta la abreviatura de los conjuntos de datos utilizados, la columna Características representa el número de variables independientes, enseguida la columna Registros que representa la cantidad de elementos contenidos en cada uno de los conjuntos de datos, por último, las columnas de No morosos y Morosos representan las solicitudes que son autorizadas y denegadas respectivamente. Se puede apreciar que ambos conjuntos de datos no están balanceados. En la Tabla A.1 del Apéndice A se describen los valores mínimos, máximos, promedio y la desviación estándar de cada una de las características contenidas en los conjuntos de datos.

Tabla 1

Elementos de los conjuntos de datos

Conjunto de datos	Características	Registros	No morosos	Morosos
GCD	24	1,000	700	300
SOCAP	23	5,237	3,482	1,755

Nota. Elaboración propia en relación con el conjunto de datos de Hofmann (1994) y la SOCAP (2022).

2.2. Preprocesamiento de datos

Los conjuntos de datos del mundo real son susceptibles de presentar diversos problemas de calidad, como son los valores nulos, distintas estructuras de datos, datos duplicados, entre otras. La calidad de los datos desempeña un papel muy importante para mejorar el rendimiento de los modelos (Priyanga & Kai, 2016). Siendo ésta una fase fundamental para realizar los ajustes necesarios, ya que se elimina el ruido, instancias y variables que no aportan valor al conjunto de datos antes de ser procesados por los modelos de aprendizaje computacional (Gemp et al., 2017).

En la Tabla A.1 se observa que los datos son dispersos y existen 11 características que tienen la desviación estándar inferior a 1 para la base de datos *GCD*, mientras que, para el conjunto de datos de la SOCAP, ocho características presentan una desviación menor a 1, por otro lado, para la clave actividad, la desviación estándar es muy grande en comparación a las demás características.

Canchen (2019) señala que la normalización (estandarización) es el procedimiento que consiste en transformar los datos de forma que todas las características se encuentren aproximadamente en la misma escala, de lo contrario las variables con mayor desviación estándar dominarán sobre las que tengan una desviación estándar más baja durante el entrenamiento de los clasificadores. La transformación de los datos es el proceso que consiste en modificar la representación de los datos de manera que estén calificados para ser la entrada de los modelos de clasificación de aprendizaje computacional. Las características pueden ser estandarizadas mediante las ecuaciones 2 y 3.

El primer método de normalización transforma los datos aplicando el método *min-max* que se describe en la ecuación 2. Este método escala todas las características del conjunto de datos entre el rango de 0 y 1.

$$X_i : \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Donde X_i, X_{min}, X_{max} representan cada uno de los valores de entrada, y los valores mínimos y máximos de las características del conjunto de datos.

Un segundo método de normalización conocido como estandarización, cambia el dominio de los datos para tener una media de 0 y una desviación estándar de 1, este proceso se realiza aplicando la ecuación 3 sobre cada columna de datos en X:

ISSN 2992-6823

$$Z: \frac{X - X_{mean}}{X_{stddev}} \tag{3}$$

Donde X_{mean} y X_{stdev} son el promedio y la desviación estándar de cada una de las características en el conjunto de datos. Siendo ésta con la cual se obtiene mejor rendimiento de los modelos utilizados.

2.3. Métricas de rendimiento

El *ML* es el estudio científico de los algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar una tarea específica sin programarse explícitamente (Wu et al., 2007). Para el desarrollo de esta investigación se han utilizado los clasificadores *SVM* (Vapnik y Cortes, 1995), *DNN* (McCulloch & Pitts, 1943), *DT* (Maimon & Rokach, 2014), *XGBoost* (Chen & Guestrin, 2015) *y RF* (Breiman, 2001), siendo estos los más utilizados en la literatura para la predicción del riesgo crediticio. Por lo tanto, es importante seleccionar los mejores parámetros para que el modelo propuesto pueda realizar la clasificación de los datos con una mayor precisión mediante distintas medidas de rendimiento.

La matriz de confusión propuesta por Kohavi & Provost (1998) es una herramienta para evaluar el desempeño de los clasificadores de aprendizaje computacional en el caso de la predicción del riesgo crediticio, esta matriz se muestra en la Tabla 2, donde cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias pertenecientes a la clase real. En otras palabras, la matriz de confusión permite identificar qué tipos y cantidad de aciertos y errores obtiene el modelo durante el proceso de aprendizaje con los datos (Barrios, 2019).

Tabla 2

Matriz de confusión

		Clase predicha	
		Buenos	Malos
Clase actual	Buenos	TP	FN
	Malos	FP	TN

Nota. Elaboración propia con base en Barrios (2019).

Para medir el rendimiento de los clasificadores utilizados para el análisis del riesgo crediticio se utilizaron las métricas: exactitud, precisión, sensibilidad, especificidad, error tipo I y II, *balanced error rate* y la correlación de *Matthews* (Matthews, 1975). Estas medidas se definen en las ecuaciones 4, 5, 6, 7, 8, 9, 10 y 11. En las ecuaciones *TP* representa los verdaderos positivos, *TN* los verdaderos negativos, *FP* los falsos positivos y *FN* los falsos negativos (Peng et al., 2011).

- Exactitud (*Accuracy*): Es la proporción de instancias clasificadas correctamente con relación al total de instancias. En otras palabras, mide la capacidad del modelo para predecir correctamente tanto casos positivos como negativos. Este valor se calcula usando la ecuación 4.

$$Accuracy: \frac{TN + TP}{TP + FP + FN + TN} \tag{4}$$

- Precisión (*Precisión*): Se refiere al porcentaje de casos positivos clasificados correctamente con relación al total de casos positivos. Se calcula con la ecuación 5.

$$Precisión: \frac{TP}{TP + FP} \tag{5}$$

ISSN 2992-6823

- Sensibilidad (Recall): Es conocida como la tasa de verdaderos positivos, siendo la porción de casos positivos que se clasificaron correctamente. Se calcula con la ecuación 6.

$$Recall: \frac{TP}{TP + FN} \tag{6}$$

- Especificidad (*Specify*): Es conocida como la tasa de verdaderos negativos, siendo la porción de casos negativos que se clasificaron correctamente. Se calcula con la ecuación 7.

$$Specificity: \frac{TN}{TN + FP} \tag{7}$$

- Correlación de *Matthews* (*MCC*, por sus siglas en inglés): Se define como un índice estadístico fiable para modelos de clasificación binaria, produciendo una puntuación alta sólo si la predicción obtuvo buenos resultados. Se calcula con la ecuación 8.

$$MCC: \frac{TP * TN \ - \ FP * FN}{\sqrt{(TP \ + \ FP) * (TP + FN) * (TN \ + FP) * (TN \ + FN)}} \tag{8}$$

- *Balanced Error Rate*: Es la media aritmética de la suma de la sensibilidad (ecuación 6) y la especificidad (ecuación 7), mediante la ecuación 9 cuando este valor se encuentra más cercano a 1 significa que el modelo está clasificando correctamente los patrones de entrada.

$$BER: 1 - \frac{Recall \ + \ Specificity}{2} \tag{9}$$

- Error tipo I: Representa margen de error presentado por el modelo al identificar una solicitud de alto riesgo como de bajo riesgo.

$$FNR: FN / (FN + TP) \tag{10}$$

- Error tipo II: Representa margen de error presentado por el modelo al identificar una solicitud de bajo riesgo como de alto riesgo.

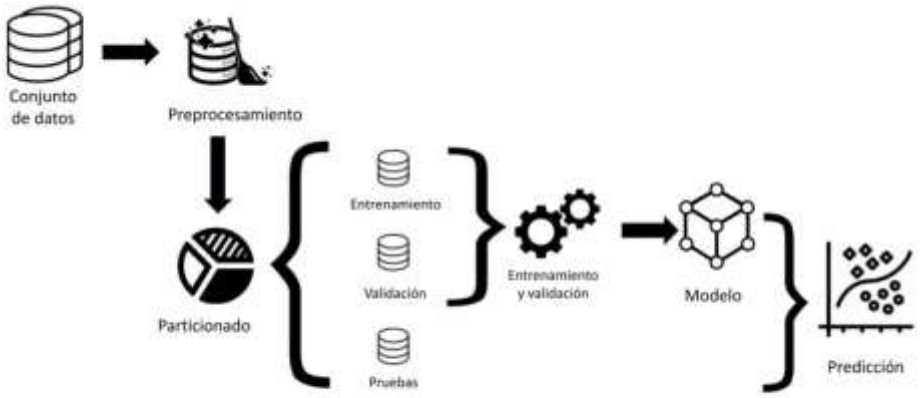
$$FPR: FP / (FP + TN) \tag{11}$$

3. Metodología

La metodología utilizada en el desarrollo se encuentra compuesta por 5 etapas (ver Figura 2).

Figura 2

Metodología del modelo propuesto.



ISSN 2992-6823

En el primer paso, se obtienen los conjuntos de datos propuestos para la identificación del riesgo crediticio. Así mismo, información relevante y características de ambos conjuntos de datos como son: edad, ocupación, género, actividad económica, estado civil, tipo de vivienda, dependientes económicos, entre otros. Los datos son proporcionados por la SOCAP en formato *CSV* y se almacenaron en una base de datos de *PostgreSQL* para facilitar el procesamiento de los datos contenidos en los archivos CSV por medio de tablas, de una manera más estructurada, desde donde se hizo la primera parte del preprocesamiento, mientras que el segundo conjunto de datos analizado se obtuvo en formato *txt* en un archivo separado por comas del repositorio *UCI Machine Learning Repository*.

El segundo paso consiste en el preprocesamiento de los datos, siendo la parte más importante la normalización de los conjuntos de datos utilizando la desviación estándar. El objetivo de este paso radica en mejorar la calidad de los datos, de tal forma que esto permita aumentar el rendimiento y la fiabilidad de los modelos de aprendizaje automático.

El tercer paso consiste en la separación de los conjuntos de datos en el conjunto de entrenamiento y pruebas de manera aleatoria en una proporción 70:30. Este paso es fundamental para evaluar y validar el rendimiento del modelo de manera objetiva y para garantizar que el modelo pueda generalizar bien los datos no vistos previamente, así el rendimiento del modelo puede ser evaluado de manera imparcial y realista.

El cuarto paso consiste en el entrenamiento de los clasificadores y selección de hiperparámetros mediante el método de búsqueda en malla, en el cual se genera un arreglo con valores cercanos a los propuestos en la literatura con la finalidad de encontrar aquellos con los cuales se alcance el mejor rendimiento de los clasificadores. Se aplicó validación cruzada en 10 pliegues durante el entrenamiento de los clasificadores para obtener estadísticas más confiables del modelo. El último paso consiste en la evaluación de los clasificadores sobre el conjunto de pruebas aplicando los hiperparámetros obtenidos en el paso anterior.

4. Resultado y discusión

El preprocesamiento de datos es el encargado de asegurar la calidad de los datos que serán analizados por los modelos propuestos. A continuación, se mencionan las técnicas de preprocesamiento utilizadas para los dos conjuntos de datos.

- **Verificar registros nulos:** Para el conjunto de datos de la SOCAP se realizó la revisión de aquellas características con valores faltantes o nulos, donde se eliminaron estos tipos de registros.
- **Validar datos con formato incorrecto:** Los registros en el conjunto de datos tienen inconsistencias o aparecen en distinto formato. Se requiere unificar criterios y formatos, depurando de esta manera el conjunto de datos.
- **Eliminar valores duplicados:** Se encontraron registros duplicados y estos valores afectan el entrenamiento de los modelos, por lo que fue necesario eliminarlos y evitar de esta manera generar ruido en el análisis.
- **Normalizar datos:** Significa comprimir o extender los valores de la variable para que estos se encuentren en un rango definido, el método utilizado para las dos bases de datos fue la estandarización por medio de la desviación estándar, la cual se define en la ecuación 3.

El conjunto de entrenamiento se particionó utilizando el método de validación cruzada en 10 pliegues para el ajuste de hiperparámetros. Finalmente, el conjunto de pruebas es utilizado para evaluar la capacidad de generalización de los clasificadores ya entrenados, estos datos nunca son presentados al clasificador durante su entrenamiento y permiten dar una medida más realista de la capacidad del modelo para generalizar ante casos nuevos.

El aprendizaje a partir de las bases de datos desbalanceadas es uno de los problemas cruciales que surgen al momento de realizar el entrenamiento de los clasificadores. La desproporción entre la cantidad de elementos de cada clase afecta el proceso de entrenamiento de los clasificadores, provocando que estos presenten una tendencia hacia la clase mayoritaria y causar la omisión de la clase minoritaria (Dablain, et al., 2023). En ambos casos de estudio, los dos conjuntos de datos se encuentran desbalanceados, ya que la proporción de los no morosos es mayor a los morosos, por tal razón es necesario aplicar una técnica de sobremuestreo utilizando el método *SMOTE* la cual consiste en sobremuestrear la clase minoritaria basándose en la regla del vecino más cercano usando la métrica de la distancia *Euclideana* (Chawla, et al., 2002).

En la Tabla 3 se observa el resultado de aplicar el sobremuestreo por medio de la técnica *SMOTE*. En la columna Balanceado se presenta el conjunto de datos balanceado y el total de registros es presentado en la columna Entrenamiento.

ISSN 2992-6823

Tabla 3

Resultado de la aplicación de sobremuestreo utilizando la técnica SMOTE.

Conjunto de datos	No balanceado	*Balanceado*	Entrenamiento
GCD	147/343	343/343	686
SOCAP	859/1,706	1,706/1,706	3,412

Nota. Elaboración propia en relación con el conjunto de datos de Hofmann (1994) y la SOCAP (2022).

4.1. Entrenamiento y validación del modelo

En esta sección se describen los parámetros utilizados por cada uno de los modelos analizados. En la Tabla 4 se presentan los hiperparámetros obtenidos durante el entrenamiento de las *SVM* con los cuales se obtiene el mejor rendimiento durante el entrenamiento.

Tabla 4

Resultado del ajuste de hiperparámetros para la SVM en ambos conjuntos de datos.

Conjunto de datos	*kernel*	*gamma*	*C*
GCD	*linear*	*1*	10
SOCAP	*poly*	*1*	0.1

Nota. Elaboración propia aplicando el método SVM para los conjuntos de datos de Hofmann (1994) y la SOCAP (2022).

Para las *DNN* se realiza la búsqueda del número de neuronas en 3 capas ocultas en la red por medio de la regla de la pirámide geométrica descrita por Grabusts & Zorins (2015) y definida por las ecuaciones 12, 13 y 14.

Primera capa oculta definida por *H1*:

$$H1 = 2 * r^3 \tag{12}$$

Segunda capa oculta definida por *H2*:

$$H2 = 2 * r^2 \tag{13}$$

Tercera capa oculta definida por *H3*:

$$H3 = 2 * r \tag{14}$$

donde $r = (número\ de\ entradas\ /\ 2)^{\frac{1}{4}}$

Partiendo de estos valores se realiza una búsqueda en malla con valores cercanos a los números de neuronas encontradas para cada capa oculta, con la finalidad de obtener la mejor arquitectura de red. Datta (2020) recomienda inicializar los pesos sinápticos de la red por medio del método de *Xavier* y las funciones de activación utilizadas son la *tanh* en las capas ocultas y la *sigmoid* en la capa de salida, siendo estas las que tienen mayor rendimiento al momento de realizar la clasificación.

En la Tabla 5 se muestran la cantidad de capas, neuronas ocultas, funciones de activación, época donde se obtiene el mejor rendimiento para las *DNN* utilizadas para cada uno de los conjuntos de datos.

ISSN 2992-6823

Tabla 5

Resultado del ajuste de parámetros del clasificador DNN para ambos conjuntos de datos.

Conjunto de datos	Función de activación	Capas Ocultas	Neuronas	Época
GCD	*tanh/sigmoid*	3	40, 25, 4,1	195
SOCAP	*tanh/sigmoid*	3	40,20,4,2	51

Nota. Elaboración propia aplicando DNN para los conjuntos de datos de Hofmann (1994) y la SOCAP (2022).

En la Tabla 6 se describen los hiperparámetros de configuración para la *DT* en los dos conjuntos de datos. Donde la calidad de división es la función para maximizar la ganancia del modelo utilizando el criterio *entropy* para la base de datos de la SOCAP y *gini* para el conjunto de datos *GDC*. Con una profundidad del árbol de 8 y la estrategia utilizada para elegir la división de cada nodo es utilizando la mejor (*best*) para elegir la mejor división aleatoria.

Tabla 6

Ajuste de parámetros de DT *para ambos conjuntos de datos*

Conjunto de datos	Calidad de división	Profundidad del árbol	Divisor
GCD	*gini*	8	*best*
SOCAP	*entropy*	8	*best*

Nota. Elaboración propia aplicando *DT* para los conjuntos de datos de Hofmann (1994) y la SOCAP (2022).

En la Tabla 7 se muestran los hiperparámetros obtenidos durante el entrenamiento del clasificador *XGBoost*, y en la Tabla 8 los correspondientes para el clasificador *Random Forest.*

Tabla 7

Ajuste de parámetros de XGBoost para ambos conjuntos de datos.

Conjunto de datos	learning_rate	max_depth	n_estimators
GCD	0.1	2	*100*
SOCAP	*0.1*	5	*60*

Nota. Elaboración propia, hiperparámetros para el clasificador *XGBoost* para los conjuntos de datos.

Tabla 8

Ajuste de parámetros de Random Forest para ambos conjuntos de datos.

Conjunto de datos	max_depth	max_features	min_samples_leaf	min_samples_split
GCD	100	3	*3*	8
SOCAP	90	3	5	8

Nota. Elaboración propia, hiperparámetros para el clasificador *Random Forest* para los conjuntos de datos.

4.2. Predicción

Se han desarrollado una cantidad importante de clasificadores de aprendizaje computacional para tratar el problema de riesgo crediticio en instituciones financieras. La Tabla 9 muestra algunos de los trabajos más recientes utilizando el conjunto de datos *GCD*, previamente descritos en la Revisión de la literatura.

ISSN 2992-6823

Tabla 9

Evaluación del rendimiento de clasificadores en trabajos relacionados con la base de datos GCD.

Artículo	**Clasificador**	***Accuracy***
Shen et al. (2021)	LSTM	80.32
Liu et al. (2021)	mg-GBDT	77.15
Lappas & Yannacopoulos (2021)	GA+KNN/GA+NB	76.16/75.69
Moscato et al. (2021)	RF	71.70
Zhang & Qiu (2020)	BP-ANN	82.53
Teles et al. (2020)	DNN	81.85
Tripathi et al. (2020)	ELM	77.92
Kuppili et al. (2019)	SELM	87.53
Dharwadkar & Pantil (2018)	DNN	76.00

Nota. Elaboración propia y del análisis de los trabajos relacionados.

Para la evaluación de los clasificadores en las bases de datos analizadas se utilizaron las medidas de rendimiento *accuracy, precision, recall, specificity, MCC y BER* los cuales son obtenidas con el conjunto de datos de pruebas, estos porcentajes se presentan en las Tablas 10 y 11.

En la Tabla 10 se muestran los resultados de aplicar los clasificadores de *SVM, DNN, DT, Random Forest y XGBoost*. Los clasificadores propuestos fueron capaces de predecir correctamente 231, 286, 191, 220 y 226 solicitudes respectivamente de un total de 300 clientes contenidos en el conjunto de pruebas. La *DNN* es el clasificador que logra obtener el mejor resultado en las medidas de rendimiento utilizadas e indican que el clasificador se encuentra realizando una buena clasificación del conjunto de datos. La arquitectura utilizada se presenta en la Tabla 5. A partir de los resultados obtenidos se concluye que el modelo de *DNN* logra una mayor precisión en la

identificación del riesgo de crédito para la base de datos *GCD* sobre los demás clasificadores y los resultados reportados en el estado del arte.

Tabla 10

Evaluación del rendimiento de clasificadores para la base de datos GCD.

Clasificador	Accuracy	Sensitivity	Specificity	ROC AUC	MCC	BER
SVM	0.7700	0.6400	0.8133	0.7024	0.4284	0.2733
DNN	**0.9533**	**0.9222**	**0.9667**	**0.9444**	**0.8889**	**0.0556**
DT	0.7133	0.5208	0.8039	0.6683	0.3306	0.3376
Random Forest	0.7333	0.6562	0.7425	0.5905	0.2686	0.3006
XGBoost	0.7533	0.6333	0.7833	0.6587	0.3637	0.2917

Nota. Elaboración propia con los modelos aplicados para el conjunto de datos de Hofmann (1994) y la SOCAP (2022).

En la Tabla 11 se muestra el rendimiento de los métodos de *SVM*, *DNN*, *DT*, *Random Forest y XGBoost*, usando el conjunto de datos de la SOCAP. Los clasificadores propuestos fueron capaces de identificar correctamente a 1369, 1377, 1409, 1431 y 1453 de un total de 1,572 elementos contenidos en el conjunto de pruebas. Siendo *XGBoost* el clasificador que tiene el mejor desempeño sobre el conjunto de pruebas en todas las métricas propuestas.

ISSN 2992-6823

Tabla 11

Evaluación del rendimiento de los clasificadores para la base de datos de créditos de la SOCAP.

Clasificador	Accuracy	Sensitivity	Specificity	ROC AUC	MCC	BER
SVM	0.8709	0.8739	0.8632	0.8709	0.7039	0.1314
DNN	0.8760	0.8559	0.9391	0.8975	0.7198	0.1025
DT	0.8963	0.8923	0.9249	0.8651	0.7636	0.1007
Random Forest	0.9103	0.8924	0.9595	0.8742	0.7985	0.0741
XGBoost	**0.9243**	**0.9127**	**0.9533**	**0.8970**	**0.8292**	**0.0670**

Nota. Elaboración propia con los modelos aplicados a la base de datos de la SOCAP (2022).

5. Conclusiones, recomendaciones y consideraciones finales

En este trabajo de investigación se han utilizado los clasificadores de aprendizaje computacional *SVM, DNN, DT, Random Forest* y *XGBoost* para analizar el riesgo crediticio. Las *DNN* ofrecen una mayor precisión cuando se aplican al conjunto de datos *GCD* para identificar el riesgo crediticio, mientras que para el conjunto de datos de la SOCAP el clasificador *XGBoost* funciona de mejor manera para el análisis del riesgo crediticio, mostrando los detalles de las características más significativas del conjunto de datos, comprobando de esta manera la hipótesis inicial.

Por lo anterior, se demuestra que al utilizar la metodología propuesta en la Sección 3, permitió alcanzar el objetivo de esta investigación al construir el modelo *DNN-XGBoost* para la clasificación del riesgo crediticio, obteniendo rendimientos superiores a 0.90 en exactitud para el nuevo conjunto de datos en comparación con otros autores (ver Tabla 9). Las características obtenidas para la SOCAP después del análisis son similares a las que se reportan para el conjunto *GCD*. Con esto podemos agregar que a pesar de que son realidades diferentes, las características del conjunto *GCD* sirvieron como base para la experimentación con la microfinanciera.

El modelo propuesto, desde una perspectiva local, servirá a la SOCAP a identificar el perfil de los socios evaluando su solvencia, lo que reducirá el riesgo de impagos, costos y tiempo de respuesta en las solicitudes de crédito, además de diseñar productos financieros adaptados a las necesidades específicas de los socios. Esto facilitará el acceso al crédito, promoviendo la inclusión financiera y el desarrollo económico local, mejorando la calidad de vida de los que integran la cooperativa. A nivel regional, se podrán identificar patrones de comportamiento en el crédito que son específicos de la región, en el estado de Oaxaca, lo que permitirá a las SOCAP ajustar sus estrategias y cumplir con su razón de ser, de impulsar el crecimiento económico. A nivel nacional, puede contribuir a la innovación y al desarrollo tecnológico en el sector financiero, mejorando la eficiencia y efectividad de las SOCAP, lo que permitirá fortalecer el sistema financiero y por ende la estabilidad económica del país.

Cabe aclarar que existen otras propuestas relacionadas con este tema de investigación, sin embargo, estas se concentran en analizar conjuntos de datos con mayor cantidad de muestras como los de bancos comerciales o empresas de carácter internacional.

Como trabajo a futuro se propone analizar el conjunto de datos con otras técnicas de aprendizaje computacional o híbridos, aplicando métodos como: la selección de características más relevantes, selección de instancias y la segmentación de datos para identificar patrones dentro de los subconjuntos formados mediante la utilización de técnicas de *clustering*, con la finalidad de mejorar el rendimiento de los clasificadores. Se sugiere dar un seguimiento de la aplicación de este modelo en otras instituciones financieras para identificar perfiles de créditos que pueden ser aceptados.

Agradecimientos

Se agradece a CONAHCYT por el financiamiento para la realización de este trabajo relacionado a la tesis “ANÁLISIS DE RIESGO CREDITICIO PARA LAS SOCAP UTILIZANDO APRENDIZAJE AUTOMÁTICO” con número de CVU 1150059 y la SOCAP con presencia en el Estado de Oaxaca con la cual se hizo un convenio para obtener los datos a analizar.

ISSN 2992-6823

Conflicto de interés

Los autores declaran que no existe conflicto de interés.

Disponibilidad de datos

Los autores utilizaron bases de datos de dominio público (disponible https://sinvestigacion.utm.mx/proyectos/IC_2024_02.html) y una base de datos privada, la cual será compartida bajo la autorización de la SOCAP.

Referencias

Barajas, A., Gutiérrez, D., Rodríguez, L & Durán, V. (2019). Indicadores para la Aprobación de Créditos Bancarios con Algoritmos Genéticos. *Aristas: Investigación Básica y Aplicada*, 7(14), 178-183. http://fcqi.tij.uabc.mx/usuarios/revistaaristas/numeros/N14/27.pdf

Barrios, J. (2019). La matriz de confusión y sus métricas. *En Big Data. Ciencia de datos, Informática Médica, Inteligencia Artificial, Machine Learning.* https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications. 39*(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, *3(*26). https://doi.org/10.3389/frai.2020.00026

Canchen, L. (2019). Preprocessing methods and pipelines of data mining: An overview. *arXiv preprint*, 1–7. https://doi.org/10.48550/arXiv.1906.08510

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://doi.org/10.1613/jair.953.

Chen, T., & Guestrin, C. (2015). Xgboost: Reliable large-scale tree boosting system. En *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* (13-17). http://learningsys.org/papers/LearningSys_2015_paper_32.pdf

Comisión Nacional Bancaria y de Valores. (CNBV) (2022). Información de Sociedades Cooperativas de Ahorro y Préstamo (SOCAP) al cierre de enero de 2022. https://www.gob.mx/cms/uploads/attachment/file/723324/Comunicado_de_Prensa_30_socaps_enero_2022.pdf

Comisión Nacional Bancaria y de Valores. (CNBV) (2017). Sociedades cooperativas de ahorro y préstamo. Administración Integral de Riesgos. https://www.gob.mx/cms/uploads/attachment/file/236226/2_Observaciones_recurrentes_en_Administraci_n_de_Riesgos.pdf

Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros. (CONDUSEF). (2021). Sociedades cooperativas de ahorro y préstamo (SOCAP). https://www.condusef.gob.mx/?p=mapa-socap&ide=1

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (9) 6390-6404. https://doi.org//10.1109/TNNLS.2021.3136503

Datta, L. (2020). A survey on activation functions and their relation with Xavier and He Normal initialization. *Neural and Evolutionary Computing*. https://doi.org/10.48550/arXiv.2004.06632

Dharwadkar, N. V., & Pantil, P. S. (2018). Customer retention and credit risk analysis using ANN, SVM and DNN. *Int. J. Society Systems Science* 10(4), 316-332. https://doi.org/10.1504/IJSSS.2018.095601

Gallardo, F., Hernández, A. & Salazar, A. (2023). Efecto de la crisis en las Sociedades Cooperativas de Ahorro y Préstamo Mexicanas. *Denarius, 1*(44), 11-50. https://doi.org/10.24275/uam/izt/dcsh/denarius/v2023n44/Gallardo

Gemp, I., Theocharous, G., & Ghavamzadeh, M. (2017). Automated Data Cleansing through Meta-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence,* 31(2), 4760-4761. https://doi.org/10.1609/aaai.v31i2.19107

Grabusts, P. & Zorins, A. (2015). The influence of hidden neurons factor on neural network training quality assurance. *Environment Technology Resources Proceedings of the International Scientific and Practical Conference*, 3, 76-81. https://doi.org/10.17770/etr2015vol3.213

ISSN 2992-6823

Guevara, C. B. & Freire, J. (2021). Modelo de clasificación de riesgo crediticio utilizando random forest en financiera del Ecuador *(Tesis, para optar el Título de Master en Sistemas de Información con Mención en Data Science).* Universidad Internacional SEK. https://repositorio.uisek.edu.ec/handle/123456789/4256

Guillén, E. & Peñafiel, L. (2017). Modelos predictor de la morosidad con variables macroeconómicas. *Revista Ciencia Unemi*, *11*(26), 13-24. https://www.redalyc.org/journal/5826/582661257002/html/

Hofmann, H. (1994). Statlog (German Credit Data) (GCD) [base de datos]. En *UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.* https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Kohavi, R., & Provost, F. (1998). Glossary of terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning*, 30, 271-274. https://doi.org/10.1023/A:1017181826899

Kuppili, V., Tripathi, D., & Edla, D. R. (2019). Credit score classification using spiking extreme learning machine. *Computational Intelligence*, *36*(2), 402-426. https://doi.org/10.1111/coin.12242

Lappas, Z., & Yannacopoulos, N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, *107*, 107391. https://doi.org/10.1016/j.asoc.2021.107391

Li, T., Kou, G., & Peng, Y. (2023). A new representation learning approach for credit data analysis. *Information Sciences*, *627*, 115-131. https://doi.org/10.1016/j.ins.2023.01.068

Liu, W., Fan, H., & Xia, M. (2021). Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Engineering Applications of Artificial Intelligence*, *97*(1), https://doi.org/10.1016/j.engappai.2020.104036

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications*, 81. World scientific.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

McCulloch, W.S., & Pitts, W (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* ***5***(4), 115-133. https://doi.org/10.1007/BF02478259

Millán, J. C., & Caicedo, E. (2018). Modelos para otorgamiento y seguimiento en la gestión de riesgo de crédito. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *25*(1), 23-41. https://doi.org/10.46661/revmetodoscuanteconempresa.2370

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, 113986. https://doi.org/10.1016/j.eswa.2020.113986

Ossa, W. & Jaramillo, V. (2021). Machine learning para la estimación del riesgo de crédito en una cartera de consumo *(Tesis, para optar el Título de Magíster en Administración Financiera)*. Universidad EAFIT. https://repository.eafit.edu.co/items/b56605a2-d3d2-4f86-9a39-cacc576c7ea9

Pandey, T. N., Mohapatra, S. K., Jagadev, A. K., & Dehuri, S. (2017). Credit Risk Analysis using Machine Learning Classifiers. *International Conference on Energy, Communication Data Analytics and Soft Computing (ICECDS), 1850-1854.* https://doi.org/10.1109/ICECDS.2017.8389769

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing 11*(2), 2906-2915. https://doi.org/10.1016/j.asoc.2010.11.028

Pérez, J., Álvaro Leitao, & García Rodríguez, J. (2023). Boundary-safe PINNs extension: Application to non-linear parabolic PDEs in counterparty credit risk. *Journal of Computational and Applied Mathematics*, 425, 115041. https://doi.org/10.1016/j.cam.2022.115041

Priyanga, C., & Kai, Q. (2016). The impact of data preprocessing on the performance of a naive bayes classifier. *En 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, 2, 618-619. https://doi.org/10.1109/COMPSAC.2016.205

Ramos, H. M. (2017). Implementación de una herramienta de análisis de riesgo de crédito basado en el modelo de rating de crédito, algoritmos genéticos y clustering jerárquico aglomerativo. *(Tesis, para optar el Título Profesional de Ingeniero de Sistemas).* Universidad Nacional Mayor de San Marcos. https://cybertesis.unmsm.edu.pe/handle/20.500.12672/7145

Rayo, S., Lara, J. & Camino D. (2010) A Credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance and Administrative Science*, *15*(28), 89-124. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2077-18862010000100005&lng=es&tlng=es

ISSN 2992-6823

Ross, Q. (1987). Statlog (Australian Credit Approval) (ACA) [base de datos]. En *UCI Machine Learning Repository*. *Irvine, CA: University of California, School of Information and Computer Science*. https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval)

Sano, Ch. (1992). Japanese Credit Screening Data Set. (JCS) [base de datos]. En *UCI Machine Learning Repository*. *Irvine, CA: University of California, School of Information and Computer Science*. https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening

Shen, F., Zhao, X., Kou, G. & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing Journal*, 98, 106852. https://doi.org/10.1016/j.asoc.2020.106852

Sun, T., & Vasarhelyi, M. A. (2018). Predicting credit card delinquencies: An application of deep neural networks. *Intelligent Systems in Accounting, Finance and Management*, *25*(4), 174-189. https://doi.org/10.1002/isaf.1437

Tavana, M., Abtahi, A.-R., Di Caprio, D., & Poortarigh, M. (2018). An artificial neural network and bayesian network model for liquidity risk assessment in banking. *Neurocomputing*, 275, 2525-2554. https://doi.org/10.1016/j.neucom.2017.11.034

Teles, G., Rodrigues, J. J. P. C., Rabê, R. A. L., & Kozlov, S. A. (2020). Artificial neural network and bayesian network models for credit risk prediction. *Journal of Artificial Intelligence and Systems*, *2*(1), 118-132. https://doi.org/10.33969/AIS.2020.21008

Trejo-García, J. C., Ríos-Bolívar, H., & Martínez-García, M. A. (2016). Análisis de la administración del riesgo crediticio en México para tarjetas de crédito. *Revista Mexicana de Economía y Finanzas* *11*(1), 103-121. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103&lng=es&tlng=es

Trejo, J. C., Ríos, H., & Almagro, F. (2016). Actualización del modelo de riesgo crediticio, una necesidad para la banca revolvente en México. *Revista Finanzas y Política Económica*, *8*(1), 17-30. http://dx.doi.org/10.14718/revfinanzpolitecon.2016.8.1.2

Tripathi, D., Edla, D. R., Kuppili, V., & Bablani, A. (2020). Evolutionary extreme learning machine with novel activation function for credit scoring. *Engineering Applications of Artificial Intelligence*. *96*(1), 103980. https://doi.org/10.1016/j.engappai.2020.103980

van Thiel, D., & van Raaij, W. F. (2019). Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era. Journal of Accounting and Finance, *19*(8), 150-170. https://doi.org/10.33423/jaf.v19i8.2622

Vapnik, V., & Cortes, C. (1995). Support-vector networks. *Machine learning*, *20*, 273-297. https://doi.org/10.1007/BF00994018

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

West, D. (2000). Neural network credit scoring models. *Computers & operations research*, *27*(11-12), 1131-1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and information systems*, 14, 1-37. https://doi.org/10.1007/s10115-007-0114-2

Yang, M., Lim, M. K., Qu, Y., Li, X., & Ni, D. (2023). Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction. *Expert Systems with Applications*, 213, 118873. https://doi.org/10.1016/j.eswa.2022.118873

Zhang, R., & Qiu, Z. (2020). Optimizing hyper-parameters of neural networks with swarm intelligence: A novel framework for credit scoring. *PLOS ONE*, 15(6), 1-35. https://doi.org/10.1371/journal.pone.0234254

Zhu, Q., Ding, W., Xiang, M., Hu, M., & Zhang, N. (2023). Loan default prediction based on Convolutional Neural Network and LightGBM. *International Journal of Data Warehousing and Mining (IJDWM)*, *19*(1), 1–16. https://www.igi-global.com/pdf.aspx?tid=315823&ptid=310110&ctid=4&oa=true&isxn=9781668479025

ISSN 2992-6823

Apéndice A

En la Tabla A.1 se describe cada una de las variables contenidas en las bases de datos Alemana y de la SOCAP. En cada una de ellas se muestran los valores mínimos, máximo, media aritmética y desviación estándar. Los datos fueron organizados considerando características similares entre ambas bases de datos. Puede notarse que características como la edad aparece en ambas bases de datos, pero en GCD la edad aparece como categórica y en la SOCAP se considera la edad como valor fijo, que posteriormente se normaliza para su uso en el entrenamiento de los clasificadores.

Tabla A.1

Análisis de los conjuntos de datos

German Credit Dataset					SOCAP				
Atributo	Min	Max	Med	Ds	Atributo	Min	Max	Med	Ds
Edad	1	2	1.15	0.36	Edad	18	87	49.13	12.79
Monto de crédito	1	5	2.10	1.58	Monto	500	500,000	195,580.13	30,338.49
Residencia	1	3	2.67	0.70	Código postal	70000	70777	70446.80	253.56
Finalidad	2	184	32.71	28.25	Finalidad	1	2	1.85	0.35
Dependientes	0	1	0.90	0.29	Dependientes económicos	0	14	2.05	2.45
Número de créditos	0	1	0.23	0.42	Créditos trabajados	0	40	8.16	6.87
Empleo	0	1	0.10	0.30	Clave actividad	1.0e+10	9.9e+10	7.1e+8	3.5e+9
Tasa de interés	1	4	2.84	1.10	Tasa normal	0	72	39.40	13.31
Cuenta de ahorro	1	5	2.10	1.58	Monto garantía	0	6	0.18	0.72
Duración	4	72	20.90	12.05	Plazo	1	55	7.35	7.11

Trabaja extranjero	0	1	0.17	0.38	Remesas	1	2	1.00	0.09
Vivienda	0	2	1.03	0.18	Tipo vivienda	1	4	2.16	0.53
Aval	19	75	35.54	11.37	Avales	0	3	0.37	0.54
Propiedad	1	4	1.40	0.57	Bien	1	5	1.10	0.64
Género & estado civil	1	4	2.35	1.05	Género	1	2	1.40	0.49
Estatus cuenta	1	4	2.57	1.25	Estado civil	1	5	2.67	1.12
Historial	0	4	2.54	1.08	Nivel académico	1	7	5.28	1.34
Planes de pago	1	2	1.40	0.49	Teléfono	0	1	0.99	0.03
Teléfono	0	1	0.04	0.19	Ingreso	0	9	2.75	3.04
					Egreso	0	9	1.43	1.77
					Tipo préstamo	1	28	4.50	5.98
					Tasa moratoria	0	72	39.40	13.31

Nota. Elaboración propia analizando los atributos de los conjuntos de datos de Hofmann (1994) y la SOCAP (2022).

ISSN 2992-6823

Apéndice B

Para el conjunto de datos *GCD* el análisis de variables mediante el método de Chi-Cuadrado se presenta en la Tabla B.1. Mientras que en la Tabla B.2 se muestra el análisis mediante el método *ANOVA*.

Tabla B.1

Análisis de variables mediante Chi - cuadrado

Característica	Valor	Existe Correlación
estatus_cuenta	1.2189E-26	Si
duracion	7.7846E-06	Si
historial_creditos	1.2792E-12	Si
finalidad	4.8687E-02	Si
monto_credito	2.7612E-07	Si
cuenta_ahorro	1.0455E-03	Si
empleo_actual	2.2238E-02	Si
genero_estadocivil	2.8584E-05	Si
residencia	1.6293E-03	Si
vivienda	1.5831E-02	Si
numero_creditos	2.8566E-03	Si
empleo	2.3490E-03	Si
trabajo_extranjero	4.4533E-03	Si
salida	1.9413E-218	Si
tasa_interes	8.6155E-01	No
aval	2.7954E-01	No
propiedad	4.4514E-01	No
edad	1.0000E+00	No
planesdepago	2.7888E-01	No
dependientes	1.0000E+00	No

telefono	7.0351E-02	No

Tabla B.2.

Análisis de variables mediante ANOVA.

Característica	Valor	Existe Correlación
estatus_cuenta	2.4417E-30	Si
duracion	6.4880E-12	Si
historial_creditos	2.4231E-13	Si
finalidad	9.8216E-07	Si
monto_credito	1.2148E-08	Si
cuenta_ahorro	2.3679E-04	Si
empleo_actual	5.2613E-03	Si
genero_estadocivil	5.9741E-06	Si
aval	3.9253E-03	Si
residencia	5.0185E-04	Si
vivienda	9.4119E-03	Si
numero_creditos	2.1575E-03	Si
empleo	1.5798E-03	Si
telefono	4.7354E-02	Si
trabajo_extranjero	3.3162E-03	Si
salida	0.0000E+00	Si
tasa_interes	9.2534E-01	No
propiedad	1.4842E-01	No
edad	9.2414E-01	No
planesdepago	2.4928E-01	No
dependientes	9.8107E-01	No

ISSN 2992-6823

En la Tabla B.3 y B.4 se describen los valores de las características del conjunto de datos de la SOCAP mediante los métodos de Chi-Cuadrada y *ANOVA.*

Tabla B.3.

Análisis de características mediante Chi-cuadrada.

Característica	Valor	Existe Correlación
edad	0.0000E+00	Si
codigopostal	1.9858E-42	Si
tipovivienda	2.0166E-07	Si
dependientes	6.1346E-91	Si
estadocivil	1.2513E-38	Si
genero	1.2681E-06	Si
claveactividad	5.4580E-67	Si
nivelacademico	3.6047E-90	Si
ingreso	2.0790E-05	Si
egreso	1.9199E-08	Si
tipoprestamo	4.2599E-84	Si
tasanormal	4.4084E-51	Si
tasamoratoria	4.4084E-51	Si
monto	4.7223E-05	Si
avales	3.0509E-14	Si
creditostrabajados	3.8745E-93	Si
bien	4.9174E-02	Si
finalidad	3.4185E-25	Si
plazo	4.5519E-13	Si
estatuscuenta	0.0000E+00	Si
telefono	3.7605E-01	No

montogarantia	1.6291E-01	No
remesas	2.4639E-01	No

Tabla B.4.

Análisis de características mediante el método ANOVA.

Característica	Valor	Existe correlación
edad	0.0000E+00	Si
tipovivienda	1.1738E-07	Si
dependientes	1.1606E-79	Si
estadocivil	1.7266E-03	Si
género	1.0671E-06	Si
claveactividad	1.8137E-36	Si
nivelacademico	1.1788E-20	Si
tipoprestamo	1.4673E-06	Si
tasanormal	3.4625E-08	Si
tasamoratoria	3.4625E-08	Si
avales	3.1785E-14	Si
creditostrabajados	7.0630E-61	Si
bien	3.9674E-02	Si
finalidad	1.2851E-25	Si
estatuscuenta	0.0000E+00	Si
codigopostal	4.0598E-01	No
telefono	2.0775E-01	No
ingreso	7.9570E-02	No
egreso	5.3462E-01	No
monto	6.5708E-01	No
montogarantia	9.1209E-01	No
remesas	1.9135E-01	No

ISSN 2992-6823

plazo	9.7739E-02	No