REVISTA MEXICANA DE ECONOMÍA Y FINANZAS Nueva Época REMEF (THE MEXICAN JOURNAL OF ECONOMICS AND FINANCE)

Revista Mexicana de Economía y Finanzas, Nueva Época

Volumen 19 Número 4, Octubre - Diciembre 2024, pp. 1-16, e868

DOI: https://doi.org/10.21919/remef.v19i4.868

(Recibido: 30/abril/2023, aceptado: 27/noviembre/2023, publicado: 27/agosto/2024)

imef fundación de investigación

Predicción del riesgo crediticio a microfinanciera usando aprendizaje computacional

Erwis Melchor Pérez - Universidad Tecnológica de la Mixteca, México
Moisés Emmanuel Ramírez Guzmán - Universidad Tecnológica de la Mixteca, México
Araceli Hernández Jiménez[1] - Universidad del Istmo, México
Agustín Santiago Alvarado - Universidad Tecnológica de la Mixteca, México

Resumen

El principal riesgo que enfrentan las Sociedades Cooperativas de Ahorro y Préstamo según la Comisión Nacional Bancaria y de Valores, es el crédito. En este artículo se aplican modelos híbridos de aprendizaje computacional para la predicción del riesgo crediticio de solicitudes de clientes pertenecientes a estas sociedades, además se describe la importancia de la selección de características y la reducción de la dimensionalidad, combinando métodos de aprendizaje no supervisado y supervisado. Los experimentos mostraron que los modelos híbridos en conjunto con técnicas de selección de características superan a los algoritmos de aprendizaje computacional de manera individual utilizando todas las características de los conjuntos de datos analizados. Los conjuntos están desbalanceados, por lo cual se utiliza el método de SMOTE para sobremuestrear la clase minoritaria y equilibrar la cantidad de elementos durante el entrenamiento. Los resultados obtenidos confirman que la combinación de métodos no supervisados y supervisados generan una mejora del 6% en el accuracy en comparación con los modelos del estado del arte y 10% en la reducción del error del tipo II para las bases de datos públicas analizadas.

Clasificación JEL: G21, C22, C44, C45, C52, C53.

Palabras clave: Instituciones Microfinancieras, Redes neuronales, Árbol de decisión, XGBoost, SMOTE.

Microfinance Credit Risk Prediction Using Computational Learning

Abstract

According to the National Banking and Securities Commission, the main risk faced by Savings and Loan Cooperative Societies is credit. This paper applies hybrid computational learning models to predict the credit risk of applications from customers belonging to these societies, and describes the importance of feature selection and dimensionality reduction, combining unsupervised and supervised learning methods. Experiments showed that hybrid models in conjunction with feature selection techniques outperform computational learning algorithms individually using all the features of the analyzed data sets. The data sets are unbalanced, so the SMOTE method is used to oversample the minority class and balance the number of features during training. The results obtained confirm that the combination of unsupervised and supervised methods generate a 6% improvement in accuracy compared to the state of the art models and 10% reduction in type II error for the analyzed public databases.

JEL Classification: G21, C22, C44, C45, C52, C53.

Keywords: Microfinance institutions, Neuronal networks, Decision tree, XGBoost, SMOTE.

[1] Autor de correspondencia. Email:
*Sin fuente de financiamiento para el desarrollo de la investigación

1. Introducción

El avance tecnológico y la creación de nuevas aplicaciones basadas en aprendizaje automático (ML por sus siglas en inglés) se han ido incorporando en el sector financiero. Donde se pretende crear modelos de riesgos crediticios con el fin de desarrollar herramientas que puedan ayudar a las microfinancieras. La selección de características tiene la finalidad de escoger las más apropiadas para la evaluación del riesgo crediticio, ayudando de esta manera a mejorar la precisión de los modelos utilizados para la toma de decisiones. Los avances tecnológicos y la transformación de nuevas aplicaciones se han incorporado a las entidades financieras (Li et al. 2022). El presente trabajo presenta la construcción de un modelo de riesgo crediticio basado en aprendizaje automático (ML) que pretende ser una herramienta que ayude a las Sociedades Cooperativas de Ahorro y Préstamo (SOCAP).

Las SOCAP son entidades que representan un segmento importante de las microfinanzas en México, las cuales tienen como objetivo principal contribuir a la inclusión financiera de la población de las comunidades en las que operan, a fin de hacerles llegar productos y servicios financieros de calidad como lo son: el crédito, ahorro e inversión que contribuyan a mejorar su situación económica (CONCAMEX, 2023), sin embargo, enfrentan distintos retos como lo son: retención de socios, detección de fraudes, segmentación de clientes y gestión de riesgos, éste último según Rivas, Cabanilla y Coello (2021), puede impactar de manera positiva en su rentabilidad siempre y cuando los riesgos crediticios se seleccionen y monitoreen de manera eficiente para mejorar las predicciones de los niveles de riesgo futuros. Por tal motivo, se ha propuesto un modelo de aprendizaje computacional que pueda brindar ayuda a la solución del problema en el sector de las microfinancieras para el análisis del riesgo crediticio con resultados de técnicas de aprendizaje supervisado como lo son las redes neuronales (NN), máquinas de soporte vectorial (SVM), *eXtreme Gradient Boosting* (*XGBoost*, por sus siglas en inglés) y árboles de decisión (DT).

El reconocimiento de patrones es una herramienta útil para el análisis y procesamiento de conjuntos de datos con muchas características, es razonable suponer que conservando todas las características se obtiene mayor conocimiento, sin embargo, es bien sabido que las características irrelevantes y redundantes pueden tener un impacto negativo en los algoritmos de aprendizaje, disminuyendo el rendimiento de los clasificadores. La selección de características permite reducir la dimensionalidad de los datos, facilitando la visualización y comprensión de estos, además suele conducir a modelos más compactos con mejor capacidad de generalización.

El resto de este documento se encuentra conformado de la siguiente manera: en la sección 2 se presenta una breve descripción de los trabajos relacionados, en la sección 3, se describe el método utilizado. Los experimentos se presentan y discuten en la sección 4. Por último, en el apartado 5 se concluye y enuncian futuras investigaciones sobre este tema.

2. Trabajos relacionados

Hand y Henley (1997) definen al crédito como aquella cantidad de dinero que una entidad financiera presta a un consumidor y éste deberá devolver con intereses en un determinado plazo. Se denomina *credit scoring* a todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a una solicitud de crédito. El riesgo calculado está en función de variables como: solvencia del solicitante, tipo de crédito, plazo, finalidad del crédito, y de otras características propias de la persona que solicita el préstamo (Medina et al. 2013).

Rayo Cantón et al. (2010) proponen que las entidades financieras dispongan de un modelo de *credit scoring* que les permita medir la probabilidad de impago de los créditos solicitados, por tanto, es necesario utilizar métodos estadísticos y modelos inteligentes para determinar dicha estimación.

Las aplicaciones de clasificación crediticia demuestran que la opinión de expertos es muy importante, así mismo la importancia y capacidad de cada una de las características o atributos de los socios (Lappas y Yannacopoulos, 2021). Por esto, las instituciones financieras tienen como objetivo lograr una identificación precisa entre una persona morosa y no morosa, generando de esta manera la utilización de varios algoritmos de ML como las NN, DT, SVM, agrupamiento y regresión lineal (RL, por sus siglas en inglés).

Chang et al. (2018) utilizan el clasificador *XGBoost* para construir un modelo que permita evaluar el riesgo crediticio para instituciones financieras. Otros autores como Lee et al. (2002) proponen el uso de modelos híbridos con la finalidad de obtener una mejora en la precisión al realizar la clasificación crediticia. Mientras que Li et al. (2022) utilizan un modelo híbrido para identificar el riesgo crediticio utilizando *XGBoost* como selector de características y las técnicas de NN, SVM y DT para obtener un modelo con mayor rendimiento al momento de realizar la clasificación. Por otra parte, Zhou et al. (2021) utilizan modelos de regresión lineal para la selección de características y obtener aquellas que mantengan la mayor varianza.

En el trabajo realizado por Machado y Karray (2022) describen la necesidad que tienen las empresas financieras para identificar el riesgo crediticio, por tal motivo proponen el uso de un algoritmo híbrido que combine los métodos de aprendizaje automático no supervisados y supervisados. Así mismo, Bao et al. (2019) proponen un modelo híbrido utilizando aprendizaje no supervisado y supervisado, evaluando los modelos individualmente y aplicando algoritmos de agrupamiento a los conjuntos de datos, resultando ser eficaz para mejorar el rendimiento de los modelos al realizar la clasificación.

3. Metodología

El método propuesto está inspirado en el modelo híbrido de Machado y Karray (2022) utilizando la base de datos de una SOCAP. En la primera etapa, se aplica la reducción de la dimensionalidad (PCA y LDA) y selección de características (LASSO). A continuación, se aplican algoritmos de aprendizaje no supervisado (*k-means* y *DBSCAN*) para agrupar los clientes en función de un conjunto de características. En la Figura 1, se puede observar gráficamente el modelo híbrido propuesto.

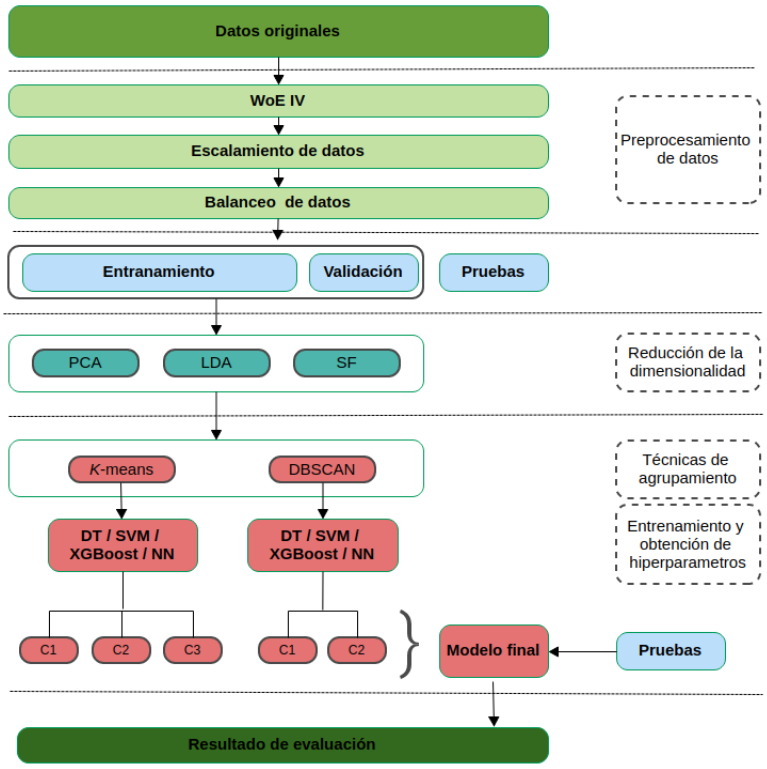


Figura 1. Método híbrido propuesto
Fuente: Elaboración propia con base en Machado y Karray (2022).

3.1 Conjunto de datos

En este estudio se utiliza el conjunto de datos proporcionado por una SOCAP, microfinanciera con presencia en el estado de Oaxaca, ésta contiene información de sus socios y se encuentra constituida por 2 clases, con 5,510 registros, de los cuales 1,683 corresponden a la clase 0 (no morosas) y 3,827 a la clase 1 (morosas). Estas clases están definidas por los días de mora que tienen los socios de la entidad al cierre de operaciones del mes de agosto del 2022. Aquellas personas que tienen más de 2 días de mora son calificadas como morosas mientras aquellas que tengan de 0 a 2 días de mora son catalogadas como no morosas. Adicional al análisis de la base de datos de la SOCAP, se muestra el resultado de procesar los 3 conjuntos de datos más utilizados por otros autores, los cuales se encuentran disponibles en *UCI Machine Learning Repository*[2]. El conjunto de datos Aleman[3] proporcionado por Hofmann (1994), el conjunto de datos Australiano[4] realizado por Ross (1987) y el conjunto de datos Japonés[5] creado por Sano (1992). La Tabla 1 describe los conjuntos de datos utilizados en la implementación de los modelos de aprendizaje computacional. Se implementa la validación cruzada para los conjuntos de datos que fueron divididos en proporción de 80% para entrenamiento y validación, mientras que el 20% para las pruebas.

[2] https://archive.ics.uci.edu/ml/index.php
[3] https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
[4] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29
[5] https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening

Tabla 1. Descripción de los conjuntos de datos utilizados.

Conjunto de datos	Total de registros	No morosas / Morosas	No. de características
Alemana	690	307/383	14
Australiana	690	307/383	15
Japonesa	1,000	700/300	20
SOCAP	5,510	1683/3827	23

Fuente: Elaboración propia con base en los conjuntos de datos de: Hofmann (1994), Ross (1987), Sano (1992) y de la SOCAP.

3.1.1 Preprocesamiento de los datos

La fase del preprocesamiento de los datos se llevó a cabo utilizando el gestor de base de datos PostgreSQL, para realizar la construcción de la base de datos de la SOCAP. La información proporcionada fue de manera anonimizada por temas de confidencialidad y se realizaron las siguientes tareas:

- Limpieza de datos faltantes.
- Unificar y eliminar registros duplicados.
- Las características de los datos en formato de texto fueron categorizados a un formato numérico.
- Se utilizó la fórmula de *Weight of Evidence (WoE*, por sus siglas en inglés*)* e *Information Value (IV*, por sus siglas en inglés*)*, para obtener el poder predictivo de las variables y la mejor combinación de grupos formados por cada uno de los atributos categóricos que se encuentran en la base de datos, como se observa en las ecuaciones:

$$WOE = ln\frac{distribucióndenomorosos}{distribucióndemorosos} \quad (1)$$

$$IV = \Sigma(\%demorosos - \%denomorosos) * WOE \quad (2)$$

donde:

- distribución de no morosos: es el porcentaje de clientes pertenecientes a la clase de no morosos en un grupo en particular.
- distribución de morosos: es el porcentaje de clientes pertenecientes a la clase de morosos en un grupo en particular.
- *ln*: logaritmo natural.

De acuerdo con Siddiqi (2012), por convención, los valores de la estadística IV, la clasificación crediticia se puede interpretar considerando la Tabla 2:

Tabla 2. Reglas relacionadas con el IV.

Valor de la información	Predicción de la variable morosos/ no morosos
< 0.02	La variable no es útil para la predicción.
0.02 - 0.1	La variable tiene un poder predictivo débil.
0.1 - 0.3	La variable tiene un poder predictivo medio.
0.3 - 0.5	La variable tiene un poder predictivo fuerte.
> 0.5	La variable tiene un poder predictivo sospechoso.

Fuente: Adaptado de (Siddiqi, 2012).

Además, se utiliza el método del valor máximo y mínimo llamado *min–max* para la normalización de los datos. Este método normaliza los valores de las características entre un rango de 0 y 1 (Machado y Karray, 2022).

Como se observa en la Tabla 1, las bases de datos se encuentran desbalanceadas, esto llega a afectar el rendimiento de los modelos de aprendizaje computacional, siendo esta característica quien origina uno de los problemas más cruciales al momento de realizar el entrenamiento de los modelos (Dablain et al. 2022). Para resolver esto, se recurre a utilizar la técnica *Synthetic Minority Oversampling TEchnique* (*SMOTE,* por sus siglas en inglés), para sobremuestrear la clase minoritaria. Produciendo nuevas instancias dentro del espacio de características ampliando la clase minoritaria, mediante la siguiente ecuación.

$$x_{new} = x_i + \left(x_i^k - x_i\right) * \delta$$

donde x_i^k es uno de los vecinos más cercanos a x_i,y δ es un valor aleatorio entre 0 y 1.

3.1.2 Reducción de la dimensionalidad

La selección de características es una tarea importante en el preprocesamiento de datos, donde el objetivo es eliminar aquellas características irrelevantes y/o redundantes del conjunto de datos analizado (Tsai et al. 2013). En el trabajo de Jia et al. (2022) definen a las bases de datos con muchas características como de alta dimensionalidad, por tanto, es conveniente transformar los conjuntos de datos de alta dimensión a un espacio de dimensión relativamente baja, conservando la mayor varianza de la información de la base de datos original.

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es un algoritmo de transformación lineal no supervisado que produce nuevas características denominadas Componentes Principales (PC), mediante la determinación de la varianza máxima de los datos. El PCA proyecta el conjunto de datos altamente dimensionales a un nuevo subespacio en el que los ejes ortogonales, o PC se consideran las direcciones de la varianza máxima de los datos. Durante la transformación, el primer PC tiene la varianza más alta, y los PC siguientes tienen varianzas decrecientes (Anowar et al. 2021). De igual forma describen al Análisis Discriminante Lineal (LDA, por sus siglas en inglés) como un algoritmo de extracción de características lineal y supervisado. LDA

identifica un nuevo espacio de características para proyectar los datos con el objetivo de maximizar la separabilidad de las clases. Por lo tanto, el número de componentes producido es menor que el número de clases *(m) - 1*. Una de las principales desventajas de utilizar LDA es que, para los problemas de clasificación binaria, sólo se tendrá una nueva característica del conjunto de datos original.

Otra técnica de selección de características es el método de regresión propuesto por Tibshirani (1996) llamado método del Operador de Selección y Contracción Mínima Absoluta (LASSO, por sus siglas en inglés), éste se puede representar mediante la reducción de la función de log-verosimilitud negativa sujeta a restricciones ponderadas, definida por la siguiente ecuación:

$$\Sigma_{i=1}^{n}\left[-Y_{i,y+12}(\beta_0 + \beta'^{X_{i,t}}) + \log(1 + \exp(\beta_0 + \beta'^{X_{i,t}}))\right]$$

sujeto a $\Sigma_{k=1}^{p}|\beta_k| \leq s$, donde n representa el número de instancias y p es la cantidad de características predictoras usadas en el modelo. LASSO elige a las características con mayor capacidad de predicción mediante la reducción hacia cero de algunos de sus coeficientes y la minimización de otros.

3.1.3 Métodos de clasificación

Para el desarrollo de este estudio el aprendizaje no supervisado se utiliza mediante la implementación del algoritmo del vecino más cercano (*K-means)* y *DBSCAN*, y las técnicas de aprendizaje supervisado DT, NN, SVM y el algoritmo XGBoost. En esta sección se presenta una revisión breve del funcionamiento de los diferentes algoritmos de aprendizaje supervisado y no supervisado que fueron utilizados. Así mismo, la descripción de los modelos híbridos los cuales utilizan la combinación de modelos de aprendizaje computacional, en Brazdil et al. (2008) describen que estos modelos pueden manejar conjuntos de datos grandes y mejoran el rendimiento de los modelos de manera individual.

Los métodos no supervisados de aprendizaje computacional no requieren de la información de la variable dependiente al momento de realizar la clasificación, como es el caso de las técnicas de agrupamiento, los cuales se describen a continuación.

El algoritmo K-means es un algoritmo capaz de agrupar un conjunto de datos de forma rápida y eficaz Géron (2022). La implementación de este método consta de los siguientes pasos:

1. Seleccionar *k* centroides aleatoriamente.
2. Etiquetar las muestras de cada grupo.
3. Actualizar los *k* centroides.
4. Actualizar las etiquetas de pertenencia a cada grupo.
5. Repetir los puntos 3 y 4, hasta que los *k* centroides converjan.

La agrupación espacial de aplicaciones con ruido basada en la densidad (*DBSCAN*, por sus siglas en inglés) agrupa las muestras que se encuentran cercanas a los centroides y los lejanos son considerados como los datos atípicos, esta implementación tiene los siguientes pasos Géron (2022):

1. Para cada muestra, el algoritmo realiza el conteo de cuántas instancias se encuentran a una pequeña distancia ϵ. Esta región es llamada vecindad ϵ de la muestra.
2. Si se obtiene un mínimo de muestras dentro de la vecindad ϵ es considerada una instancia central.
3. Todas las muestras cercanas a la vecindad de un centroide pertenecen al mismo grupo.
4. Cualquier muestra que se encuentra lejos de la vecindad es considerada como un dato atípico.

Mientras los algoritmos no supervisados no necesitan el tipo de clase al que pertenece cada registro, para los algoritmos supervisados es necesario tener dicha información al momento de realizar la tarea de clasificación como en los casos mencionados a continuación.

El tipo de NN más utilizado es el perceptrón multicapa (MLP, por sus siglas en inglés), el cual está conformado por una capa de entrada, una o más capas ocultas y una capa de salida. Las NN han sido utilizadas para problemas de clasificación, entre ellos la predicción del riesgo de crédito (Machado y Karray, 2022). Las redes neuronales reciben en la capa de entrada las características de cada uno de las muestras, las cuales son procesadas por las capas ocultas hasta llegar a la capa de salida, dicha capa es la encargada de presentar la predicción generada por la red, basada en los pesos obtenidos en el proceso de entrenamiento de la red (Bishop y Nasrabadi, 2008).

La NN propuesta es de 3 capas definida por medio de la regla de la pirámide geométrica descrita por Grabusts y Zorins (2015). Partiendo de estos valores se realiza una búsqueda en malla con valores cercanos a los números de neuronas encontradas para cada capa oculta. Los pesos sinápticos de la red fueron inicializados por medio del método de *Xavier* (Datta, 2020). Las funciones de activación utilizadas fueron *tanh* en la primera y segunda capa oculta, mientras que en la capa de salida fue *sigmoid*.

Los DT son un modelo de aprendizaje supervisado donde los datos son particionados continuamente de acuerdo a ciertos parámetros evaluados hasta formar el árbol (De Ville, 2013). Los datos obtenidos del árbol pueden ser explicados basándose en los nodos de decisión y sus hojas; las hojas son las decisiones finales y los nodos de decisión son los puntos donde los datos son separados. La principal característica de los DT son las divisiones de manera recursiva de un campo objetivo de datos en función a los valores de entrada para crear particiones y subconjuntos de datos en cualquier nivel del árbol. Las métricas utilizadas para maximizar la calidad de división de los árboles son la impureza de gini, la ganancia de información (entropía) y la reducción de la varianza (Tangirala, 2020).

El algoritmo *eXtreme Gradient Boosting* (*XGBoost*) está basado en el algoritmo *gradient boosting tree* (Chen et al. 2015). El algoritmo XGBoost se encuentra basado en la teoría de la clasificación y el árbol de regresión (Qiu et al. 2021). Además, la función objetivo evita que el modelo presente sobreentrenamiento. $D = \{x_i, y_i\}$ representa un conjunto de datos que contiene n ejemplos y m características, y los resultados de predicción están conformados por las siguientes ecuaciones:

$$\widehat{y_i}\Sigma_{k=1}^{k} f_k(x_i), f_k \epsilon \varphi$$

$$\varphi = \{f(x) = w_s(x)\}(s: R^m \rightarrow T, w_s \epsilon R^T)$$

donde $\widehat{y_i}$, representa la etiqueta de predicción, x_i representa una de las muestras y $f_k(x_i)$ es la muestra dada, φ simboliza el árbol de regresión, $f(x)$ y w representa el peso de las hojas y número de hojas, respectivamente.

El modelo de las SVM se entrena en base al ajuste de los parámetros, con el objetivo de crear un límite de decisión entre dos clases que pueda permitir la predicción de etiquetas a partir de uno o más vectores de características (Wang y Hu, 2005). La función *kernel* permite a la SVM determinar la forma del hiperplano y el límite de decisión de un conjunto de datos, proyectando los datos de un espacio de baja dimensión a un espacio de dimensión superior (Patle y Chouhan, 2013). Los tipos de *kernel* utilizados en las SVM son el lineal, polinomial, radial y *sigmoide*.

3.1.4 Medidas de rendimiento

En esta sección se analizan los criterios utilizados para evaluar los modelos de clasificación. El *Accuracy*, área bajo la curva (*AUC*, por sus siglas en inglés) error tipo I, error tipo II, *Balanced Error Rate* (*BER*, por sus siglas en inglés), Coeficiente de Correlación de Matthews (*MCC*, por sus siglas en inglés) estos dos últimos adecuados para para conjunto de datos desbalanceados (Chicco et al. 2021). En este estudio, los componentes principales de la matriz de confusión son: Verdadero Positivo (TP) y Falso Positivo (FP) que representan casos no morosos clasificados de manera correcta y erróneamente. Los Verdaderos Negativos (TN) y Falso Negativo (FN) representan los morosos clasificados correcta y erróneamente.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$AUC = \frac{\frac{TP}{TP + FP} + \frac{TN}{TP + FN}}{2}$$

$$ErrortipoI = \frac{FN}{TP + FN}$$

$$ErrortipoII = \frac{FP}{FP + TN}$$

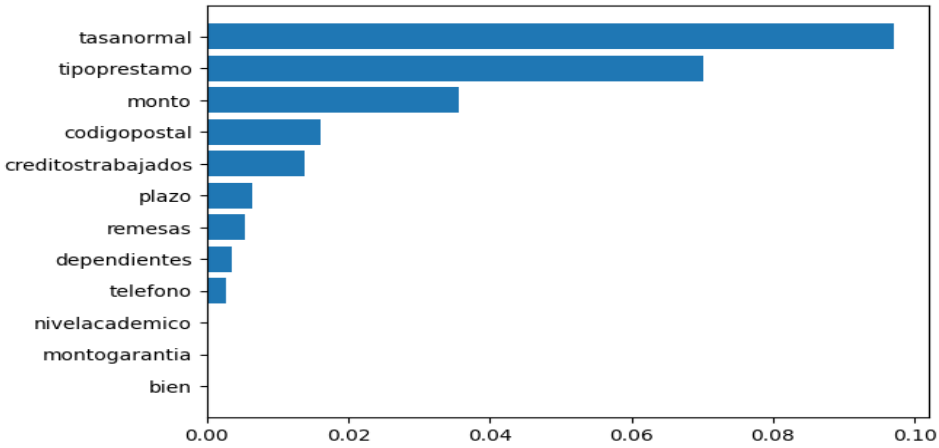
$$MCC = \frac{TP * TN - FP * FN}{\sqrt{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}\,(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}}$$

$$BER = \frac{\frac{FP}{FP + TP} + \frac{FN}{FN + TN}}{2}$$

4. Experimentación y resultados

En las tablas 3 y 4 se presenta la evaluación de los clasificadores con las 3 bases de datos libres utilizadas en este experimento. En la sección 3 se muestra el método propuesto. Los resultados que obtuvieron el primer lugar son presentados en negritas, además la optimización de los hiperparámetros de los modelos se realiza por medio de la búsqueda en malla y la validación cruzada. En la tabla 5 y 6 se muestran los resultados obtenidos con la selección de características en conjunto de los algoritmos de agrupamiento *k-means* y *DBSCAN* en conjunto de los modelos de aprendizaje computacional con la base de datos de la SOCAP.

En las instituciones financieras las características más importantes utilizadas por los expertos para la identificación de los socios morosos y no morosos son: la tasa de crédito, ingresos, egresos, monto del crédito, plazo, garantías, actividad económica y la zona de residencia. Dichas características coinciden con el resultado de utilizar el algoritmo de regresión *LASSO* como método de selección de características, como se observa en la Gráfica 1. *LASSO* retorna el valor de los coeficientes, donde aquellos con valor mayor a 0 son las características más significativas:



Gráfica 1. Listado de características más significativas obtenidas con métodos de aprendizaje computacional para la base de datos de la SOCAP.
Fuente: Elaboración propia salida del método *LASSO.*

Tabla 3. Rendimiento de los clasificadores con las tres bases de datos libres utilizando FS y *k-means*.

Base de datos	Modelo	Accuracy	AUC	Error tipo I	Error tipo II	BER	MCC
Australiana	DT	**0.9058**	0.9069	0.1000	**0.0649**	**0.0931**	**0.8088**
	NN	0.9058	0.9052	**0.0897**	0.1000	0.0949	0.8088
	SVM	0.8623	0.8602	0.1184	0.1613	0.1399	0.7214
	XGBoost	0.8623	**0.9396**	0.0972	0.1818	0.1395	0.7252

	Artículo	0.8797	0.9396	0.0749	0.0977		
Alemana	DT	**0.7750**	**0.7320**	**0.1724**	**0.3636**	**0.2680**	0.4521
	NN	0.7350	0.6880	0.1778	0.4462	0.3120	0.3844
	SVM	0.7300	0.6736	0.2410	0.4118	0.3264	**0.2847**
	XGBoost	0.6950	0.6496	0.1938	0.5070	0.3504	0.3124
	Artículo	0.9979	0.9997	0.0014	0.0029		
Japonesa	DT	0.8986	0.8985	0.1013	0.1017	0.1015	0.7940
	NN	**0.9203**	**0.9187**	**0.0658**	**0.0968**	**0.0813**	**0.8388**
	SVM	0.8143	0.8113	0.1467	0.2308	0.1888	0.6262
	XGBoost	0.8478	0.8471	0.1000	0.2059	0.1530	0.6988
	Artículo	0.8841	0.9406	0.0749	0.0699		

Fuente: Elaboración propia y comparación con los resultados obtenidos en (Zhou et al. 2021)

Tabla 4. Rendimiento de los clasificadores con las tres bases de datos libres utilizando FS y *DBSCAN*.

Base de datos	Modelo	Accuracy	AUC	Error tipo I	Error tipo II	BER	MCC
Australiana	DT	**0.9124**	**0.9179**	0.1098	**0.0545**	**0.0822**	**0.8243**
	NN	0.8478	0.8455	0.1316	0.1774	0.1545	0.6921
	SVM	0.8478	0.8471	**0.1000**	0.1818	0.1530	0.6988
	XGBoost	0.8188	0.8190	0.1750	0.1897	0.1810	0.6315
	Artículo	0.8797	0.9396	0.0749	0.0977		
Alemana	DT	0.7350	0.6810	0.2041	0.5000	0.3191	0.3486
	NN	0.7550	0.7151	**0.1473**	0.4225	0.2849	0.4492
	SVM	0.7350	0.6800	0.2298	0.4103	0.3201	0.3112
	XGBoost	**0.7840**	**0.7440**	0.1608	**0.3509**	**0.2559**	**0.4810**
	Artículo	0.9979	0.9997	0.0014	0.0029		
Japonesa	DT	0.8986	0.9070	0.1294	0.0566	0.0930	0.7972
	NN	**0.9493**	**0.9545**	0.0732	**0.0179**	**0.0456**	**0.8988**
	SVM	0.8768	0.8801	**0.0455**	0.1944	0.1200	0.7645

	XGBoost	0.8551	0.8571	0.0746	0.2113	0.1430	0.7187
	Artículo	0.8841	0.9406	0.0749	0.0699		

Fuente: Elaboración propia y comparación con los resultados obtenidos en (Zhou et al. 2021)

Los algoritmos de aprendizaje computacional implementados se llevaron a cabo en Python 3.9, con un entorno virtual usando las bibliotecas Scikit-learn, TensorFlow, Pandas y NumPy. Para obtener el mejor modelo se evalúan los hiperparámetros de cada uno de los algoritmos implementados utilizando la búsqueda en malla y validación cruzada para cada una de las bases de datos. Los resultados obtenidos demuestran que la selección de características y el sobremuestreo en bases de datos desbalanceadas juegan un papel importante en el rendimiento de los algoritmos y se comportan de distinta manera en la base de datos analizada. Así mismo, al combinar modelos de aprendizaje computacional como las técnicas de selección de características, sobremuestreo y *clustering* pueden incrementar el rendimiento de los clasificadores e identificar aquellos registros con características similares y comportamientos de los distintos solicitantes de créditos.

Combinando la selección de características y el método de *clustering k-means* obteniendo 3 grupos para el caso de la base de datos Australiana y Alemana, los árboles de decisión tienen el mejor rendimiento en cuanto al *accuracy* y el tipo de error II. Mientras que con la base de datos Japonesa el mejor *accuracy* es obtenido con las redes neuronales. Para la selección de características y el algoritmo *DBSCAN* obteniendo 2 grupos, para la base de datos Australiana el mejor rendimiento en cuanto al *accuracy* y tipo de error II es obtenido con los árboles de decisión, para la Alemana es con el algoritmo *XGBoost* mientras que el mejor rendimiento para la base de datos Japonesa es obtenido con las redes neuronales. De esta manera se deduce que la metodología propuesta tiene distinto comportamiento con cada uno de los algoritmos planteados dependiendo de la base de datos utilizada.

Para el caso de la base de datos de la SOCAP, el mejor rendimiento es obtenido con el algoritmo *XGBoost* con el método de *clustering k-mean* y *DBSCAN*, siendo éste un algoritmo potente y robusto para poder generalizar correctamente los solicitantes de créditos en la microfinanciera. El objetivo principal es reducir a lo mínimo posible el tipo de error II, que corresponde a aquellos clientes a quienes se les autoriza un crédito y no debería de ser asignado. La tabla 5 muestra el rendimiento de la configuración del experimento al aplicar el algoritmo de agrupamiento *k-means* para generar 3 grupos y después aplicar los algoritmos DT, NN, SVM y XGBoost. Así mismo, la tabla 6 muestra el resultado de aplicar el algoritmo DBSCAN, que genera 2 grupos y posteriormente, se aplican los mismos algoritmos de ML.

Tabla 5. Rendimiento de los clasificadores con la base de datos de la SOCAP utilizando FS y *k-means*.

Base de datos	Cluster	Accuracy	AUC	Error tipo I	Error tipo II	BER	MCC
k-means - DT	1	0.9890	0.9894	0.0097	0.0115	0.0106	0.9732
	2	0.9900	0.9902	0.0090	0.0105	0.0098	0.9787

	3	0.9886	0.9883	0.0087	0.0146	0.0117	0.9770
	Promedio	0.9875	0.9893	0.0087	0.0146	0.0321	0.9763
k-means - NN	1	0.9945	0.9961	0.0000	0.0077	0.0039	0.9866
	2	0.9834	0.9814	0.0264	0.0106	0.0186	0.9645
	3	0.9957	0.9954	0.0043	0.0049	0.0046	0.9908
	Promedio	0.9912	0.9909	0.0102	0.0077	0.0090	0.9806
k-means - SVM	1	0.9863	0.9819	0.0283	0.0078	0.0181	0.9667
	2	0.9867	0.9857	0.0179	0.0106	0.0143	0.9716
	3	0.9977	0.9975	0.0000	0.0049	0.0025	0.9954
	Promedio	0.9902	0.9883	0.0154	0.0077	0.0116	0.9779
k-means - XGBoost	1	0.9890	0.9866	0.0190	0.0077	0.0134	0.9232
	2	0.9967	0.9956	0.0088	0.0000	0.0044	0.9929
	3	0.9954	0.9954	0.0043	0.0049	0.0046	0.9908
	Promedio	0.9937	0.9925	0.0107	0.0042	0.0074	0.9856

Fuente: Elaboración propia con base en el conjunto de datos de la SOCAP.

Tabla 6. Rendimiento de los clasificadores con la base de datos de la SOCAP utilizando FS y *DBSCAN*.

Base de datos	Cluster	Accuracy	AUC	Error tipo I	Error tipo II	BER	MCC
DBSCAN - DT	1	0.9733	0.9800	**0.0000**	0.0400	0.0200	0.9428
	2	**0.9932**	**0.9928**	0.0111	**0.0340**	**0.0072**	**0.9862**
	Promedio	0.9833	0.9864	0.0056	0.0217	0.0136	0.9645
DBSCAN - NN	1	0.9200	0.9093	0.1379	0.0435	0.0907	0.8305
	2	**0.9981**	**0.9983**	**0.0000**	**0.0034**	**0.0017**	**0.9960**
	Promedio	0.9591	0.9538	0.0690	0.0235	0.0462	0.9133
DBSCAN - SVM	1	0.9600	0.9537	0.0714	0.0370	0.0464	0.9143
	2	**0.9951**	**0.9949**	**0.0067**	**0.0035**	**0.0051**	**0.9901**
	Promedio	0.9776	0.9743	0.0391	0.0203	0.0257	0.9522
DBSCAN - XGBoost	1	0.9467	0.9375	0.1034	0.0217	0.0626	0.8875

	2	**1.0000**	**1.0000**	**0.0000**	**0.0000**	**0.0000**	**1.0000**
	Promedio	0.9734	0.9687	0.0517	0.0109	0.0313	0.9438

Fuente: Elaboración propia con base en el conjunto de datos de la SOCAP.

Los resultados obtenidos confirman que el método de selección de características *LASSO* y las técnicas de *clustering* analizadas y combinadas con los DT y NN muestran una mejora considerable en relación a ocupar únicamente los clasificadores. Estos dos clasificadores ofrecen una mejora del *accuracy* de 0.90±0.05 con respecto a otros autores, y en cuanto al error tipo II 0.1±0.05 para los conjuntos Australiano, Alemán y Japonés. Sin embargo, el comportamiento es distinto en cada uno de los conjuntos de datos, pero a pesar de esto, los resultados son superiores a los reportados en el estado del arte. Mientras que con la base de datos de la SOCAP el mejor rendimiento es obtenido con el algoritmo *XGBoost*, reduciendo en 0.01±0.04 con otros algoritmos de *ML*. Estos resultados se logran por medio del algoritmo de *clustering* basado en la densidad de la información.

5. Conclusiones

En este estudio se presentaron los rendimientos obtenidos de los algoritmos de ML para clasificación, métodos de reducción de dimensionalidad y selección de características en tres bases de datos de dominio público y la base de datos de la SOCAP. Al aplicar la metodología propuesta se incrementa el rendimiento de los algoritmos al realizar la predicción de las personas morosas y no morosas, además de minimizar el error de los falsos positivos en comparación con los resultados reportados en el estado del arte.

Las técnicas actuales de aprendizaje computacional podrían determinar la combinación adecuada de las técnicas de selección de características y clasificadores. Además de implementar y evaluar la metodología propuesta con las bases de datos utilizadas.

La precisión del clasificador en las instituciones de crédito es importante para la salud financiera, por tanto, puede generar ganancias importantes y minimizar las pérdidas. Los resultados obtenidos indican que la combinación de las características definidas por el experto y reforzada por las técnicas de selección de aprendizaje computacional pueden mejorar la precisión del clasificador y minimizar el riesgo en el proceso de toma de decisiones. Por lo tanto, el enfoque combinado podría utilizarse en la detección de fraudes y lavado de dinero. Además, en futuras investigaciones, se puede mejorar la interpretabilidad de los resultados conforme al éxito y fracaso de los modelos implementados, analizando otras técnicas de *clustering* y selección de instancias con la finalidad de identificar aquellos registros que ingresan ruido o tienen información duplicada en el conjunto de datos, permitiendo agilizar el entrenamiento de los algoritmos de aprendizaje computacional, conservando en lo máximo posible la calidad de los datos.

Referencias

[1] Anowar, F., Sadaoui, S., y Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

[2] Bao, W., Lianju, N., y Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301-315. https://doi.org/10.1016/j.eswa.2019.02.033

[3] Bishop, C. M., y Nasrabadi, N. M. (2008). *Pattern recognition and machine learning* (4th ed., Vol. 4). Springer.

[4] Brazdil, P., Carrier, C. G., Soares, C., y Vilalta, R. (2008). *Metalearning: Applications to data mining*. Springer Science y Business Media.

[5] Chang, Y.-C., Chang, K.-H., y Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029

[6] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., y Zhou, T. (2015). XGBoost: extreme gradient boosting. *package version 0.4-2*, *1*(4), 1-4.

[7] Chicco, D., Tótsch, N., y Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, *14*(13), 1-22. https://doi.org/10.1186/s13040-021-00244-z

[8] CONCAMEX. (2023). *CONFEDERACIÓN DE COOPERATIVAS DE AHORRO Y PRÉSTAMO DE MÉXICO*. Retrieved Marzo 10, 2023, from https://www.concamex.coop/es/

[9] Dablain, D., Krawczyk, B., y Chawla, N. V. (2022). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 1-15. https://doi.org//10.1109/TNNLS.2021.3136503

[10] 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

[11] De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448-455. https://doi.org/10.1002/wics.1278

[12] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

[13] Grabusts, P., y Zorins, A. (2015). Proceedings of the International Scientific and Practical Conference. En *ENVIRONMENT. TECHNOLOGIES. RESOURCES* (Vol. 3, 76-81).

[14] Hand, D. J., y Henley, W. E. (1997). Statistical Classication Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

[15] 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)

[16] Jia, W., Sun, M., Lian, J., y Hou, S. (2022). Feature dimensionality reduction: a review. *Complex y Intelligent Systems*, *8*(3), 2663--2693. https://doi.org/10.1007/s40747-021-00637-x

[17] Lappas, P. Z., y Yannacopoulos, A. 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

[18] Lee, T.-S., Chiu, C.-C., Lu, C.-J., y Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications*, *23*(3), 245-254. https://doi.org/10.1016/S0957-4174(02)00044-1

[19] Li, Y., Stasinakis, C., y Meng Yeo, W. (2022). A Hybrid XGBoost-MLP Model for Credit Risk Assessment on Digital Supply Chain Finance. *Forecasting*, *4*(1), 184-208. https://doi.org/10.3390/forecast4010011

[20] Machado, M. R., y Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, *200*, 116889.

[21] Medina, P., María, R., Selva, M., y Luisa, M. (2013). Análisis del credit scoring. *RAE-Revista de Administração de Empresas.*, *53*(3), 303-315.

[22] Patle, A., y Chouhan, D. S. (2013). SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)* (pp. 1-9). https://doi.org/10.1109/ICAdTE.2013.6524743

[23] Qiu, Y., Zhou, J., Khandelwal, M., Yang, H., Yang, P., y Li, C. (2021). Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engineering with Computers*, *28*, 4145–4162. https://doi.org/10.1007/s00366-021-01393-9

[24] Rayo Cantón, S., Lara Rubio, J., y Camino Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science*, *15*(28), 89-124.

[25] Rivas, M.C., Cabanilla, G., y Coello, M.G. (2021). El impacto del riesgo crediticio en rentabilidad de cooperativas de ahorro y crédito ecuatorianas. Universidad y Sociedad, 13 (S3), 459-466. https://rus.ucf.edu.cu/index.php/rus/article/view/2505

[26] 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)

[27] Sano, C. (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

[28] Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley y Sons.

[29] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

[30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

[31] Tsai, C.-F., Eberle, W., y Chu, C.-Y. (2013). Genetic algorithms in feature and instance selection. *Knowledge-Based Systems*, *39*, 0950-7051. https://doi.org/10.1016/j.knosys.2012.11.005

[32] Wang, H., y Hu, D. (2005). Comparison of SVM and LS-SVM for Regression. In *2005 International Conference on Neural Networks and Brain* (1st ed., pp. 279-283). https://doi.org//10.1109/ICNNB.2005.1614615

[33] Zhou, Y., Shamsu Uddin, M., Habib, T., Chi, G., y Yuan, K. (2021). Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istraživanja*, *34*(1), 3064-3091. https://doi.org/10.1080/1331677X.2020.1867213