

ARTÍCULO ORIGINAL

Algoritmos de balanceo probabilístico: desarrollo y evaluación de un sistema informático para clasificación binaria

Probabilistic Balancing Algorithms: Development and Evaluation of a Software System for Binary Classification

Ireimis de las Mercedes Leguen de Varona

ireimis.leguen@reduc.edu.cu • <https://orcid.org/0000-0002-1886-7644>

Julio Madera Quintana

julio.madera@reduc.edu.cu • <https://orcid.org/0000-0001-5551-690X>

UNIVERSIDAD DE CAMAGÜEY “IGNACIO AGRAMONTE LOYNAZ”, CAMAGÜEY, CUBA

Alfredo Simon-Cuevas

asimon@ceis.cujae.edu.cu • <https://orcid.org/0000-0002-6776-9434>

UNIVERSIDAD TECNOLÓGICA DE LA HABANA “JOSÉ ANTONIO ECHEVERRÍA”, CUJAE, CUBA

Marcos Antonio Rodríguez Guerra

marcos.rodriguez@reduc.edu.cu • <https://orcid.org/0009-0001-0661-1932>

UNIVERSIDAD DE CAMAGÜEY “IGNACIO AGRAMONTE LOYNAZ”, CAMAGÜEY, CUBA

Recibido: 2026-02-13 • Aceptado: 2026-03-26

RESUMEN

El desbalance de clases es un problema habitual en tareas de clasificación supervisada, donde la clase minoritaria se encuentra representada por una proporción significativamente menor de instancias. Esta situación compromete la capacidad de los modelos para reconocer correctamente los casos más relevantes, como en la clasificación de texto, detección de fraudes o diagnósticos médicos. Aunque existen múltiples algoritmos de sobremuestreo que abordan este reto, las soluciones probabilísticas que se proponen en este trabajo: SMOTE-COV-LW, SMOTE-RL y SMOTE-EN, presentan un enfoque innovador basado en modelos probabilísticos. Sin embargo, su uso ha estado limitado a entornos de programación, lo que restringe su aplicación por parte de usuarios no especializados. Para facilitar su acceso, se diseñó un sistema informático que permite aplicar estos algoritmos de sobremuestreo probabilísticos para conjuntos de datos de clasificación binaria de forma gráfica y sin necesidad de conocimientos técnicos avanzados. El sistema fue desarrollado con tecnologías

como Python y PyQt6, e incluye herramientas para importar bases de conocimientos de clasificación binarias, aplicar balanceo y exportar resultados. Para evaluar el rendimiento de los algoritmos probabilísticos propuestos, se realizó una comparación con técnicas clásicas ampliamente utilizadas (SMOTE, Borderline-SMOTE, SMOTE-RSB y ADASYN). La evaluación se realizó sobre múltiples conjuntos de datos de clasificación binaria desbalanceados, utilizando como clasificadores a C4.5, MLP, KNN, Random Forest y SVM, y como métricas de desempeño el AUC y el F1-Score. Los resultados mostraron que los algoritmos probabilísticos alcanzaron resultados similares o incluso superiores en varios escenarios, lo que evidencia su competitividad frente a los métodos tradicionales.

Palabras clave: Aprendizaje Automatizado; Desbalance de Clases; Sobremuestreo Probabilístico; Modelos Probabilísticos; Sistema Informático.

ABSTRACT

Class imbalance is a common problem in supervised classification tasks, where the minority class is represented by a significantly smaller proportion of instances. This situation compromises the model's ability to correctly recognize the most relevant cases, such as in text classification, fraud detection, or medical diagnoses. While multiple oversampling algorithms address this challenge, the probabilistic solutions proposed in this work—SMOTE-COV-LW, SMOTE-RL, and SMOTE-EN—present an innovative approach based on probabilistic models. However, their use has been limited to programming environments, which restricts their application by non-specialist users. To facilitate access, a software system was designed to apply these probabilistic oversampling algorithms to binary classification datasets in a graphical manner, without the need for advanced technical knowledge. The system was developed using technologies such as Python and PyQt6 and includes tools for importing binary classification datasets, applying balancing techniques, and exporting results. To evaluate the performance of the proposed probabilistic algorithms, a comparison was made with widely used classical techniques (SMOTE, Borderline-SMOTE, SMOTE-RSB, and ADASYN). The evaluation was conducted on multiple imbalanced binary classification datasets, using C4.5, MLP, KNN, Random Forest, and SVM as classifiers, and AUC and F1-Score as performance metrics. The results showed that the probabilistic algorithms achieved similar or even superior results in several scenarios, demonstrating their competitiveness compared to traditional methods.

Keywords: Machine Learning; Class Imbalance; Probabilistic Oversampling; Probabilistic Models; Software System.

INTRODUCCIÓN

La inteligencia artificial y el aprendizaje automático son herramientas clave para extraer conocimiento en dominios críticos como el diagnóstico médico y la detección de fraudes (Soofi & Awan, 2017; Zhang et al., 2021). No obstante, su eficacia se ve limitada por el desbalance de clases, fenómeno donde la clase de interés está subrepresentada, provocando que los modelos favorezcan a la clase mayoritaria y pierdan utilidad práctica (Ramentol et al., 2012a; Krawczyk, 2016).

Se puede definir que un conjunto de datos es balanceado si tiene aproximadamente igual porcentaje de ejemplos en los conceptos a clasificar, es decir, si la distribución de ejemplos por clases es uniforme; de lo contrario, es no balanceado. Para medir el grado de desbalance de un problema se define la razón de desbalance (IR, del inglés Imbalance Ratio) (Fernández, García, del Jesus, & Herrera, 2008):

$$IR = \frac{C_+}{C_-} \quad (1)$$

Donde:

C_+ es el número de instancias que pertenecen a la clase mayoritaria.

C_- es el número de instancias que pertenecen a la clase minoritaria.

Por consiguiente, un conjunto de datos es no balanceado cuando presenta una marcada diferencia entre los ejemplos de las clases. Esta diferencia provoca que los algoritmos de clasificación tengan gran exactitud para calcular modelos sobre la clase más representada (mayoritaria), pero pobre exactitud predictiva sobre los elementos de la clase menos representada (minoritaria). Esto resulta en que, al final del proceso, las clases con mayor número de instancias quedan muy bien clasificadas mientras que las minoritarias tienden a estar mal clasificadas, debido a que los clasificadores intentan reducir el error global sin tener en cuenta la distribución de los datos. En los conjuntos no balanceados, el conocimiento original suele etiquetarse como rarezas o ruido, enfocándose exclusivamente en las medidas globales (Chawla, Lazarevic, Hall, & Bowyer, 2003; Fernández et al., 2018). El problema con los conjuntos no balanceados no es únicamente la desproporción en la representatividad, sino también el alto solapamiento entre las clases (López et al., 2013; Das et al., 2023).

Para mitigar este problema, la técnica más extendida es SMOTE (Synthetic Minority Over-sampling Technique), que genera ejemplos sintéticos mediante interpolación lineal (Chawla et al., 2002). A lo largo de los años, han surgido numerosas variantes que buscan superar sus limitaciones (Fernández et al., 2018). Sin embargo, SMOTE y sus variantes —como Borderline-SMOTE (Han et al., 2005), SMOTE-RSB* (Ramentol et al., 2012a) y ADASYN (He et al., 2008)— presentan una limitación estructural: asumen independencia entre atributos, ignorando las correlaciones intrínsecas de los datos reales (Fernández et al., 2018; Alghamdi et al., 2022).

Como alternativa, el uso de modelos gráficos probabilísticos permite modelar explícitamente estas dependencias (Madera, 2008; Murphy, 2022). Bajo este paradigma, se evalúan tres algoritmos propuestos:

SMOTE-COV-LW: emplea el estimador de Ledoit-Wolf (Ledoit & Wolf, 2022) para generar muestras que preserven la estructura de covarianza. Este tipo de enfoques basados en estimación robusta de covarianza ha mostrado ventajas en escenarios de alta dimensionalidad (Bodnar et al., 2021).

SMOTE-RL: utiliza regresión Lasso (Tibshirani, 1996; Hastie, Tibshirani & Wainwright, 2019) para la regularización y selección de variables durante la síntesis.

SMOTE-EN: aplica Redes Elásticas (Elastic Net) (Zou & Hastie, 2005) para manejar la multicolinealidad, un enfoque que ha demostrado ser eficaz en problemas con variables correlacionadas (Algarnal & Lee, 2019).

A pesar de su potencial, estas técnicas avanzadas suelen estar restringidas a entornos de programación, lo que dificulta su adopción por usuarios no especializados. Por ello, este artículo aborda una doble necesidad: realizar una evaluación experimental exhaustiva frente a métodos clásicos y presentar un sistema informático intuitivo (desarrollado en Python y PyQt6) que facilite el acceso a estas herramientas. La validación se realiza mediante cinco clasificadores (C4.5, MLP, KNN, Random Forest y SVM) y las métricas AUC y F1-Score, demostrando la competitividad de los métodos propuestos y la utilidad de la herramienta desarrollada.

METODOLOGÍA

Este estudio se caracteriza por ser de tipo experimental y comparativo, con un enfoque cuantitativo. El diseño de la investigación se estructuró en dos fases principales: (1) el desarrollo e implementación de tres nuevos algoritmos probabilísticos de balanceo de datos (SMOTE-COV-LW, SMOTE-RL y SMOTE-EN) junto con un sistema informático para su uso accesible; y (2) la evaluación experimental rigurosa de estos algoritmos frente a técnicas clásicas de sobremuestreo.

La investigación es reproducible, ya que se basa en conjuntos de datos públicos, métodos algorítmicos descritos formalmente, métricas de evaluación estandarizadas y un protocolo experimental claramente definido. El enfoque metodológico combina el desarrollo de software (para la creación del sistema) con la validación empírica (para la evaluación de los algoritmos).

Sobremuestreo Probabilístico

Para generar muestras sintéticas semánticamente coherentes, se emplean tres métodos basados en Modelos Gráficos Probabilísticos (MGP), que son estructuras de grafos donde los nodos representan variables aleatorias y las aristas representan dependencias condicionales entre ellas. Estos grafos proporcionan una representación compacta de las distribuciones de probabilidad (Madera, 2008). Entre los MGP, los Modelos Gráficos Gaussianos se destacan como modelos de interacción para distribuciones normales multivariadas (J. Richardson Ibáñez, 2017). Técnicas como la Matriz de Covarianza, la Regresión Lasso y la Red Elástica se utilizan para estimar estos Modelos Gráficos Gaussianos.

Matriz de Covarianza

La matriz de covarianza Σ mide la relación lineal entre las variables aleatorias de un vector. Se define a partir de las varianzas σ_i^2 y las covarianzas σ_{ij} entre las variables i y j . Para mejorar la estimación de esta matriz, especialmente en espacios de alta dimensión con pocas muestras, se emplea el estimador de contracción de Ledoit-Wolf. Este método combina la matriz de covarianza muestral S con una matriz de estructura simple F (como la matriz identidad), controlada por un parámetro de contracción δ que minimiza el error cuadrático medio:

$S^* = (1 - \delta)S + \delta F$ (2). Esta regularización previene el sobreajuste al reducir la matriz hacia una estructura más simple (Ledoit & Wolf, 2022).

Regresión Lasso

La regresión Lasso es una técnica de regularización útil para datos de alta dimensión, ya que minimiza la suma de los errores residuales cuadrados aplicando una penalización $L1$ sobre el valor absoluto de los coeficientes β_j :

$$\min_{\beta} \left(\sum_{i=1}^n \left(y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda \sum_j |\beta_j| \right) \quad (3).$$

El parámetro λ controla la penalización. A medida que aumenta, algunos coeficientes se reducen exactamente a cero, lo que previene el sobreajuste y realiza una selección automática de variables (Lu & Yin, 2022).

Red Elástica (Elastic Net)

La Red Elástica combina las propiedades de las penalizaciones $L1$ (Lasso) y $L2$ (Ridge), superando las limitaciones individuales de ambas al permitir la selección de grupos de variables correlacionadas:

$$\min_{\beta} \left(\sum_{i=1}^n \left(y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j \beta_j^2 \right) \quad (4).$$

Donde λ_1 permite la selección de variables anulando coeficientes, y λ_2 reduce la colinealidad contrayendo los coeficientes sin eliminarlos completamente.

Marco general del pseudocódigo de las propuestas de sobremuestreo probabilístico

Entrada: Conjunto original D , Índice de desbalance objetivo IR_{obj} , Parámetros de regularización.

Salida: Conjunto balanceado D_{bal} .

Identificar clase minoritaria C_{min} y extraer sus instancias X_{min} .

Estandarizar atributos de X_{min} .

Entrenar Modelo de Estimación (según variante):

Si COV-LW: Estimar matriz de covarianza Σ (Ledoit-Wolf) y obtener su descomposición de Cholesky L .

Calcular vector de medias μ .

Si RL / EN: Entrenar modelo de regresión con penalización Lasso (L_1) o Elastic Net ($L_1 + L_2$).

Calcular parámetros de control (límites, rangos o percentiles de X_{min}).

Mientras $IR(D) > IR_{obj}$ hacer:

Generar vector de ruido aleatorio $\epsilon \sim N(0, I)$.

Generar instancia sintética (x_{syn}):

Si COV-LW: $x_{syn} = \mu + \epsilon \cdot L^T$.

Si RL o EN: $x_{syn} = \text{Modelo}(\epsilon)$.

Aplicar transformación inversa de la estandarización.

Ajuste de Rango (Post-procesamiento):

Si COV-LW: Aplicar límites físicos (IR) o permitir expansión probabilística (FR).

Si RL: Ajustar a límites observados de los atributos.

Si EN: Realizar ajuste basado en percentiles de los datos originales.

Añadir x_{sym} con etiqueta C_{min} al conjunto D_{bal} .

Fin Mientras

Devolver D_{bal} .

Evaluación experimental

Se utilizaron siete bases de datos binarias (Tabla 1) con Índices de Desbalance (IR) entre 1.82 y 5.46. El desempeño de las propuestas se comparó con cuatro métodos clásicos (SMOTE, Borderline-SMOTE, SMOTE-RSB y ADASYN) empleando cinco clasificadores (C4.5, MLP, KNN, Random Forest y SVM). Se aplicó validación cruzada (k=10) evaluando AUC y F1-Score.

Tabla 1. Propiedades de los conjuntos de datos seleccionados para la experimentación

Bases de datos	Número de Casos	Número de Atributos	IR
ecoli2	336	7	5.46
glass-0-1-2-3_vs_4-5-6	214	9	3.2
glass1	214	9	1.82
iris	150	4	2
newthyroid2	215	5	5.14
pima	768	8	1.87
vehicle3	846	18	2.99

El objetivo de este experimento es comparar los resultados obtenidos con los algoritmos de sobremuestreo probabilísticos frente a los métodos clásicos, para determinar si son competitivos o superiores.

Tests de validación estadísticos

Para garantizar la rigurosidad del análisis sin asumir normalidad ni homogeneidad de varianzas, se emplearon pruebas no paramétricas. Se calculó el ranking de los algoritmos utilizando la prueba de Friedman y su variante ajustada de Iman-Davenport para detectar diferencias globales significativas. Posteriormente, se seleccionó el algoritmo con mejor desempeño como método de control y se aplicó la prueba post-hoc de Holm para compararlo con el resto, ajustando los p-valores secuencialmente para controlar el error tipo I.

Diseño del sistema informático

Para materializar el aporte teórico de esta investigación, se desarrolló BalanceAI, una herramienta de escritorio dirigida a investigadores y estudiantes que facilita el tratamiento del desbalance de clases. El sistema fue implementado en Python 3.9 utilizando la biblioteca PyQt6 para el diseño de su interfaz gráfica, adoptando una arquitectura altamente modular.

El diseño del flujo de trabajo permite al usuario, sin necesidad de conocimientos avanzados en programación, realizar el proceso completo desde una única interfaz: carga de conjunto de datos, selección y configuración de parámetros para 7 algoritmos de balanceo (incluyendo los tres métodos probabilísticos propuestos), entrenamiento y evaluación con los 5 clasificadores integrados, y finalmente, la visualización y exportación tanto de los datos balanceados como de los reportes de rendimiento.

RESULTADOS Y DISCUSIÓN

Se evaluó la eficacia de las propuestas SMOTE-COV-LW (FR e IR), SMOTE-RL y SMOTE-EN frente a los algoritmos del estado del arte (SMOTE, Borderline-SMOTE, ADASYN y SMOTE-RSB). La Tabla 2 consolida el desempeño de los cinco clasificadores base sobre los siete conjuntos de datos.

En términos generales, las tres propuestas demostraron una consistencia superior. SMOTE-RL y SMOTE-EN alcanzaron los valores de F1-Score más altos en la mayoría de los escenarios (ej. 0.97 en ecoli2 con C4.5), validando que el filtrado de ruido tras el sobremuestreo optimiza la precisión frente al SMOTE tradicional. Por su parte, SMOTE-COV-LW destacó por preservar la estructura estadística del conjunto de datos; en el caso crítico de pima (C4.5), la variante FR superó al estado del arte con un AUC de 0.87.

Al emplear Random Forest, las tres propuestas dominaron la comparativa con valores de AUC 0.91 y F1-Score 0.81. Asimismo, en SVM, los métodos propuestos exhibieron una notable resiliencia ante el colapso de métricas: mientras que SMOTE-RSB cayó a un F1-Score de 0.07 en pima, SMOTE-RL y SMOTE-COV-LW mantuvieron el desempeño en 0.63 y 0.68 respectivamente. Esto confirma que la regularización de Ledoit-Wolf y el filtrado avanzado reducen efectivamente el solapamiento entre clases.

Tabla 2. Resultados consolidados de AUC y F1-Score (Formato: AUC / F1-Score).

Clasificador	Base de datos	Original	SMOTE	Borderline	ADASYN	SMOTE-RSB	SMOTE	SMOTECOV-LW (IR)	SMOTE-RL	SMOTE-EN
							COV-LW (FR)			
C4.5	ecoli2	0.83 / 0.96	0.95 / 0.93	0.96 / 0.95	0.95 / 0.94	0.89 / 0.96	0.93 / 0.90	0.93 / 0.92	0.96 / 0.97	0.95 / 0.96
	glass-0-1...	0.86 / 0.96	0.96 / 0.95	0.96 / 0.95	0.95 / 0.94	0.94 / 0.96	0.94 / 0.93	0.96 / 0.96	0.96 / 0.96	0.96 / 0.96
	glass1	0.72 / 0.81	0.84 / 0.81	0.81 / 0.80	0.74 / 0.78	0.67 / 0.81	0.85 / 0.84	0.83 / 0.83	0.80 / 0.81	0.82 / 0.81
	iris	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.98	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99
	newthyroid2	0.94 / 0.97	0.98 / 0.98	0.99 / 0.99	0.98 / 0.98	0.98 / 0.98	0.99 / 0.98	0.99 / 0.99	0.98 / 0.98	0.98 / 0.98
	pima	0.75 / 0.61	0.81 / 0.80	0.80 / 0.80	0.75 / 0.74	0.75 / 0.60	0.87 / 0.80	0.85 / 0.79	0.85 / 0.77	0.85 / 0.80
	vehicle3	0.76 / 0.86	0.84 / 0.82	0.85 / 0.81	0.83 / 0.83	0.76 / 0.86	0.88 / 0.84	0.84 / 0.81	0.88 / 0.85	0.88 / 0.85
MLP	ecoli2	0.96 / 0.97	0.97 / 0.95	0.98 / 0.97	0.97 / 0.94	0.97 / 0.97	0.93 / 0.88	0.98 / 0.94	0.98 / 0.96	0.99 / 0.96
	glass-0-1...	0.97 / 0.96	0.99 / 0.96	0.98 / 0.97	0.97 / 0.97	0.98 / 0.95	0.97 / 0.93	0.98 / 0.93	0.99 / 0.95	0.99 / 0.95
	glass1	0.70 / 0.78	0.84 / 0.75	0.85 / 0.77	0.85 / 0.79	0.77 / 0.78	0.84 / 0.78	0.87 / 0.78	0.83 / 0.77	0.84 / 0.77

Clasificador	Base de datos	SMOTE								
		Original	SMOTE	Borderline	ADASYN	SMOTE- RSB	SMOTE- COV- LW (FR)	SMOTECOV- LW (IR)	SMOTE- RL	SMOTE- EN
KNN	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	newthyroid2	1.00 / 0.99	1.00 / 0.99	1.00 / 0.99	1.00 / 0.99	1.00 / 0.99	1.00 / 0.99	1.00 / 0.97	1.00 / 0.99	1.00 / 0.99
	pima	0.80 / 0.63	0.84 / 0.78	0.83 / 0.76	0.80 / 0.76	0.82 / 0.66	0.90 / 0.79	0.86 / 0.77	0.89 / 0.82	0.89 / 0.81
	vehicle3	0.89 / 0.90	0.94 / 0.88	0.94 / 0.87	0.93 / 0.87	0.89 / 0.90	0.91 / 0.83	0.91 / 0.82	0.95 / 0.88	0.94 / 0.88
	ecoli2	0.91 / 0.96	0.96 / 0.96	0.98 / 0.96	0.94 / 0.94	0.94 / 0.97	0.87 / 0.88	0.94 / 0.94	0.98 / 0.97	0.97 / 0.97
	glass-0-1...	0.92 / 0.96	0.97 / 0.97	0.98 / 0.98	0.98 / 0.98	0.96 / 0.97	0.93 / 0.93	0.96 / 0.96	0.98 / 0.96	0.96 / 0.96
	glass1	0.81 / 0.87	0.88 / 0.87	0.88 / 0.87	0.88 / 0.88	0.80 / 0.85	0.82 / 0.83	0.83 / 0.83	0.88 / 0.84	0.84 / 0.84
	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	newthyroid2	0.98 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	0.96 / 0.96	1.00 / 0.99	0.99 / 0.99
	pima	0.66 / 0.55	0.81 / 0.82	0.80 / 0.81	0.80 / 0.81	0.67 / 0.56	0.76 / 0.76	0.73 / 0.71	0.82 / 0.77	0.77 / 0.77
	vehicle3	0.66 / 0.84	0.87 / 0.85	0.87 / 0.85	0.85 / 0.83	0.66 / 0.84	0.79 / 0.80	0.73 / 0.76	0.88 / 0.83	0.83 / 0.83
	ecoli2	0.97 / 0.98	0.99 / 0.97	0.99 / 0.97	0.99 / 0.96	0.98 / 0.98	0.97 / 0.92	0.99 / 0.97	0.99 / 0.97	0.99 / 0.97
	glass-0-1...	0.99 / 0.95	0.99 / 0.96	1.00 / 0.96	1.00 / 0.97	0.99 / 0.96	0.99 / 0.95	0.99 / 0.96	0.99 / 0.96	0.99 / 0.96
	glass1	0.90 / 0.90	0.95 / 0.90	0.95 / 0.88	0.95 / 0.89	0.91 / 0.89	0.93 / 0.88	0.93 / 0.86	0.94 / 0.89	0.94 / 0.87
Random Forest	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	newthyroid2	1.00 / 0.98	1.00 / 0.99	1.00 / 1.00	1.00 / 0.99	1.00 / 0.99	1.00 / 0.99	1.00 / 0.98	1.00 / 0.99	1.00 / 0.99
	pima	0.82 / 0.63	0.90 / 0.84	0.89 / 0.83	0.88 / 0.81	0.83 / 0.63	0.91 / 0.81	0.90 / 0.81	0.91 / 0.81	0.91 / 0.82
	vehicle3	0.87 / 0.87	0.96 / 0.88	0.96 / 0.87	0.96 / 0.88	0.87 / 0.87	0.96 / 0.86	0.94 / 0.86	0.96 / 0.87	0.96 / 0.87
	ecoli2	0.50 / 0.92	0.90 / 0.90	0.88 / 0.86	0.83 / 0.83	0.78 / 0.94	0.68 / 0.52	0.91 / 0.91	0.94 / 0.93	0.94 / 0.93
	glass-0-1...	0.93 / 0.96	0.96 / 0.96	0.96 / 0.96	0.94 / 0.94	0.95 / 0.96	0.95 / 0.95	0.97 / 0.96	0.96 / 0.96	0.96 / 0.96
	glass1	0.71 / 0.83	0.74 / 0.70	0.74 / 0.66	0.73 / 0.68	0.73 / 0.83	0.81 / 0.82	0.75 / 0.73	0.71 / 0.64	0.70 / 0.62
	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	newthyroid2	0.65 / 0.94	0.98 / 0.98	0.97 / 0.97	0.96 / 0.96	0.97 / 0.98	0.90 / 0.88	0.90 / 0.88	0.94 / 0.94	0.94 / 0.95
	pima	0.50 / -	0.70 / 0.56	0.69 / 0.55	0.65 / 0.46	0.50 / 0.07	0.54 / 0.68	0.54 / 0.68	0.73 / 0.63	0.65 / 0.45
	vehicle3	0.50 / 0.86	0.77 / 0.82	0.77 / 0.82	0.76 / 0.81	0.50 / 0.86	0.54 / 0.16	0.54 / 0.16	0.83 / 0.86	0.79 / 0.83
	ecoli2	0.50 / 0.92	0.90 / 0.90	0.88 / 0.86	0.83 / 0.83	0.78 / 0.94	0.68 / 0.52	0.91 / 0.91	0.94 / 0.93	0.94 / 0.93
	glass-0-1...	0.93 / 0.96	0.96 / 0.96	0.96 / 0.96	0.94 / 0.94	0.95 / 0.96	0.95 / 0.95	0.97 / 0.96	0.96 / 0.96	0.96 / 0.96
	glass1	0.71 / 0.83	0.74 / 0.70	0.74 / 0.66	0.73 / 0.68	0.73 / 0.83	0.81 / 0.82	0.75 / 0.73	0.71 / 0.64	0.70 / 0.62
	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
SVM	newthyroid2	0.65 / 0.94	0.98 / 0.98	0.97 / 0.97	0.96 / 0.96	0.97 / 0.98	0.90 / 0.88	0.90 / 0.88	0.94 / 0.94	0.94 / 0.95
	pima	0.50 / -	0.70 / 0.56	0.69 / 0.55	0.65 / 0.46	0.50 / 0.07	0.54 / 0.68	0.54 / 0.68	0.73 / 0.63	0.65 / 0.45
	vehicle3	0.50 / 0.86	0.77 / 0.82	0.77 / 0.82	0.76 / 0.81	0.50 / 0.86	0.54 / 0.16	0.54 / 0.16	0.83 / 0.86	0.79 / 0.83
	ecoli2	0.50 / 0.92	0.90 / 0.90	0.88 / 0.86	0.83 / 0.83	0.78 / 0.94	0.68 / 0.52	0.91 / 0.91	0.94 / 0.93	0.94 / 0.93
	glass-0-1...	0.93 / 0.96	0.96 / 0.96	0.96 / 0.96	0.94 / 0.94	0.95 / 0.96	0.95 / 0.95	0.97 / 0.96	0.96 / 0.96	0.96 / 0.96
	glass1	0.71 / 0.83	0.74 / 0.70	0.74 / 0.66	0.73 / 0.68	0.73 / 0.83	0.81 / 0.82	0.75 / 0.73	0.71 / 0.64	0.70 / 0.62
	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
	newthyroid2	0.65 / 0.94	0.98 / 0.98	0.97 / 0.97	0.96 / 0.96	0.97 / 0.98	0.90 / 0.88	0.90 / 0.88	0.94 / 0.94	0.94 / 0.95
	pima	0.50 / -	0.70 / 0.56	0.69 / 0.55	0.65 / 0.46	0.50 / 0.07	0.54 / 0.68	0.54 / 0.68	0.73 / 0.63	0.65 / 0.45
	vehicle3	0.50 / 0.86	0.77 / 0.82	0.77 / 0.82	0.76 / 0.81	0.50 / 0.86	0.54 / 0.16	0.54 / 0.16	0.83 / 0.86	0.79 / 0.83
	ecoli2	0.50 / 0.92	0.90 / 0.90	0.88 / 0.86	0.83 / 0.83	0.78 / 0.94	0.68 / 0.52	0.91 / 0.91	0.94 / 0.93	0.94 / 0.93
	glass-0-1...	0.93 / 0.96	0.96 / 0.96	0.96 / 0.96	0.94 / 0.94	0.95 / 0.96	0.95 / 0.95	0.97 / 0.96	0.96 / 0.96	0.96 / 0.96
	glass1	0.71 / 0.83	0.74 / 0.70	0.74 / 0.66	0.73 / 0.68	0.73 / 0.83	0.81 / 0.82	0.75 / 0.73	0.71 / 0.64	0.70 / 0.62
	iris	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00

Validación estadística de los resultados

Para validar rigurosamente los resultados experimentales expuestos en la sección anterior, se aplicaron pruebas estadísticas no paramétricas, recomendadas para la evaluación de algoritmos de Machine Learning sobre múltiples conjuntos de datos. Se utilizó el test de Friedman para establecer un ranking de rendimiento y el test post-hoc de Holm ($\alpha = 0.05$) para identificar si existían diferencias estadísticamente significativas respecto al algoritmo mejor posicionado (método de control). Con el objetivo de condensar el análisis de los cinco experimentos, la Tabla 3 resume los hallazgos para las métricas de AUC y F1-Score, indicando el algoritmo de control en cada caso y las hipótesis nulas rechazadas.

Tabla 3. Resumen de la validación estadística (Test de Friedman y Holm, $\alpha = 0.05$). Se indica el algoritmo con mejor ranking y aquellos superados significativamente.

Clasificador	Métrica	Algoritmo de Control (1er Lugar)	Algoritmos superados significativamente (Rechazados por Holm)
C4.5	AUC	SMOTE-COV-LW (FR)	Original
	F1-Score	SMOTE-EN*	Ninguno (Todos son estadísticamente equivalentes)
MLP	AUC	SMOTE-RL*	Original, SMOTE-RSB, COV-LW (IR), COV-LW (FR)
	F1-Score	Borderline-SMOTE	COV-LW (IR)
KNN	AUC	SMOTE-EN*	Original
	F1-Score	Original	Ninguno (Todos son estadísticamente equivalentes)
Random F.	AUC	Borderline-SMOTE	Ninguno (Todos son estadísticamente equivalentes)
	F1-Score	SMOTE	Ninguno (Todos son estadísticamente equivalentes)
SVM	AUC	SMOTE	Original
	F1-Score	SMOTE-RSB	Ninguno (Todos son estadísticamente equivalentes)

La validación estadística confirma que el enfoque probabilístico propuesto es altamente competitivo frente al estado del arte. Destaca el desempeño de SMOTE-COV-LW (Fuera de Rango) con el clasificador C4.5, posicionándose como el mejor método global en la métrica AUC.

En la mayoría de los escenarios (Random Forest, KNN y SVM), las pruebas de Holm demostraron equivalencia estadística entre la generación por matriz de covarianza y los algoritmos clásicos basados en k-vecinos. Este hallazgo es positivo, ya que valida una metodología que evita las limitaciones geométricas tradicionales. Las únicas excepciones fueron los perceptrones multicapa (MLP), donde los métodos clásicos mantuvieron una ligera superioridad.

Balance AI: Un sistema de balanceo y clasificación

Como respuesta a la necesidad de accesibilidad para investigadores no especializados, se desarrolló BalanceAI, una herramienta portable basada en Python y PyQt6. A diferencia de bibliotecas que requieren conocimientos de programación, este sistema centraliza el flujo de trabajo —desde la detección del desbalance hasta la evaluación con cinco clasificadores (C4.5, MLP, KNN, Random Forest y SVM)— en una interfaz gráfica intuitiva. Su factor diferenciador es la integración nativa de los algoritmos propuestos: SMOTE-COV-LW, SMOTE-RL y SMOTE-EN.

La arquitectura del sistema (Figura 1) es modular y escalable, organizada en cuatro componentes principales:

- DatasetManagement: Carga, validación estadística y exportación de datos (.csv, .arff).
- Balancing: Ejecución de algoritmos de sobremuestreo y visualización de distribuciones.
- Classification: Entrenamiento de modelos y cálculo de métricas (AUC y F1-Score).
- AlgorithmSelection: Configuración dinámica de parámetros de balanceo y clasificación.

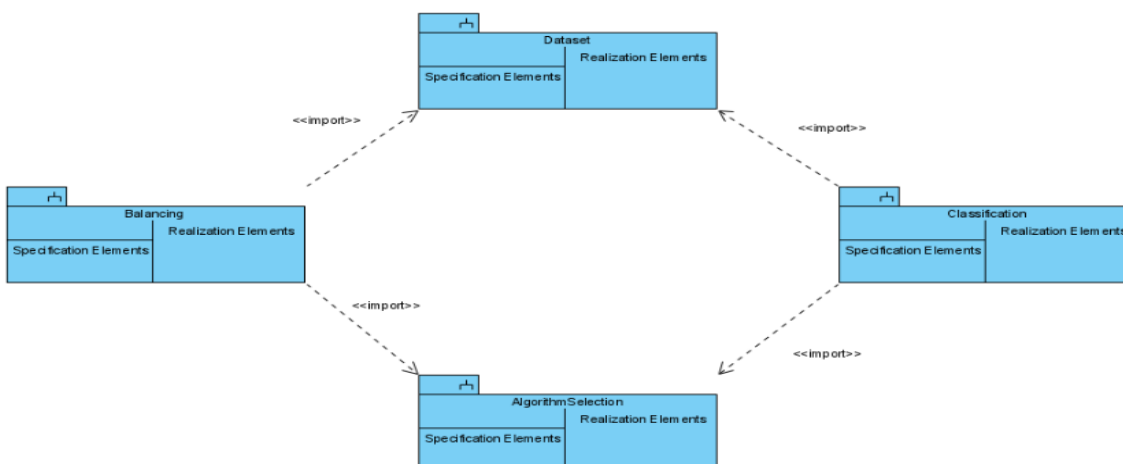


Figura 1. Diagrama de subsistemas.

La interfaz gráfica del usuario (GUI) fue diseñada priorizando la usabilidad. La Figura 2 muestra los entornos principales de trabajo, donde el usuario puede interactuar con los módulos descritos de manera fluida, desde la configuración del balanceo hasta el análisis visual de los resultados de clasificación.

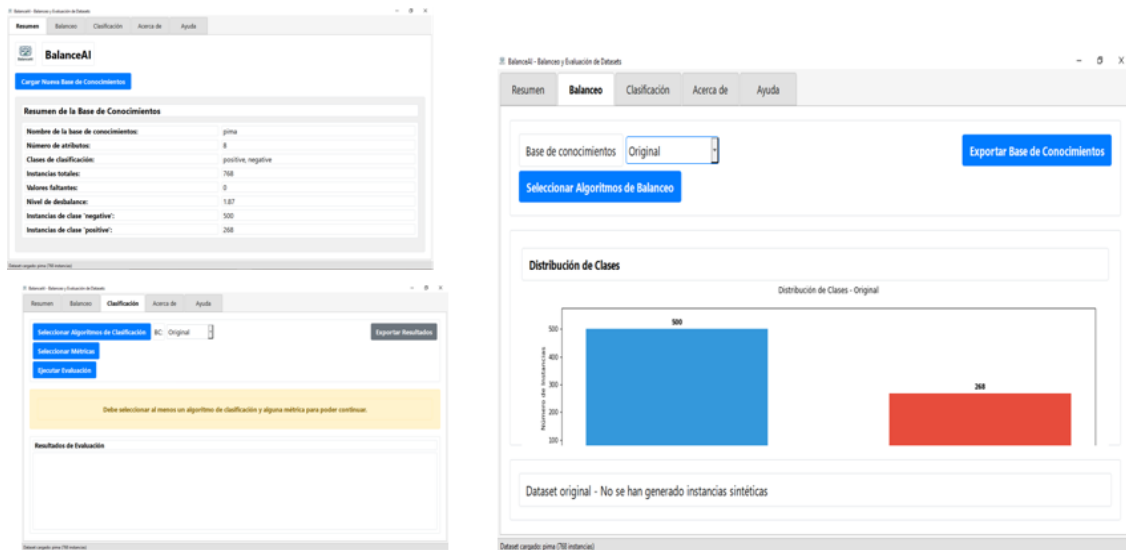


Figura 2. Interfaces principales del sistema BalanceAI (A: Configuración y Balanceo; B: Evaluación y Resultados).

CONCLUSIONES

La presente investigación abordó el desafío del desbalance de clases en tareas de clasificación binaria mediante una estrategia dual que combina la innovación algorítmica con el desarrollo de software aplicado. A partir de los resultados obtenidos, se derivan las siguientes conclusiones:

El desarrollo y la evaluación de los algoritmos SMOTE-COV-LW, SMOTE-RL y SMOTE-EN demostraron que la utilización de Modelos Gráficos Gaussianos, matrices de covarianza y técnicas de regularización (Lasso y Elastic Net) permite generar instancias sintéticas de mayor calidad semántica. Las rigurosas pruebas estadísticas no paramétricas (Friedman y Holm) confirmaron que estos enfoques propuestos superan o compiten de manera estadísticamente significativa con los métodos clásicos de sobremuestreo, mejorando métricas clave como el AUC y el F1-Score en diversos clasificadores.

La materialización de estos modelos matemáticos en la herramienta informática BalanceAI representa un aporte práctico significativo para la comunidad académica. El sistema logra democratizar el acceso a técnicas avanzadas de inteligencia artificial, ofreciendo a investigadores y estudiantes una plataforma modular, portable y de fácil uso que no requiere conocimientos avanzados de programación para ejecutar flujos complejos de balanceo y evaluación.

Como líneas de continuidad para esta investigación, se proyecta la evaluación de los algoritmos probabilísticos propuestos en conjuntos de datos multiclase y en espacios de alta dimensionalidad (como el procesamiento de lenguaje natural). Asimismo, desde el punto de vista del software, se planea dotar a BalanceAI de una arquitectura basada en plugins que permita la integración colaborativa de nuevos métodos de preprocesamiento y modelos de aprendizaje profundo (Deep Learning).

AGRADECIMIENTOS

Deseamos expresar nuestro más sincero agradecimiento al Programa Nacional de Ciencia y Tecnología PN223LH004: Automática, Robótica e Inteligencia Artificial, del Ministerio de Ciencia, Tecnología y Medio Ambiente de Cuba, por el apoyo brindado a este trabajo en el marco del proyecto PN223LH004-038: Contribuciones teóricas a la IA en la gestión de problemas de datos complejos.

REFERENCIAS

Alghamdi, M., Alghamdi, M., & Al-Barakati, A. (2022). A systematic review on oversampling techniques for imbalanced data classification. *IEEE Access*, *10*, 12458-12478.

Algama, Z. Y., & Lee, M. H. (2019). A new adaptive elastic net for high-dimensional data. *Journal of Statistical Computation and Simulation*, *89*(9), 1689-1702.

Bodnar, T., Okhrin, Y., & Parolya, N. (2021). Optimal shrinkage covariance matrix estimation in high-dimensional problems. *Journal of Multivariate Analysis*, *185*, Article 104767. <https://doi.org/10.1016/j.jmva.2021.104767>

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.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 107-119). Springer.

Das, S., Datta, S., & Chaudhuri, B. B. (2023). Handling class overlap and imbalance to detect rare events. *Pattern Recognition*, *133*, Article 109018.

Fernández, A., García, S., del Jesus, M. J., & Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule-based classification systems in the framework of imbalanced data sets. *Fuzzy Sets and Systems*, 159(18), 2378-2398.

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863-905.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

Google Cloud. (s.f.). ¿Qué es la inteligencia artificial o IA? Recuperado el 22 de junio de 2025, de <https://cloud.google.com/learn/what-is-artificial-intelligence>.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing* (pp. 878-887). Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2019). *Statistical learning with sparsity: The Lasso and generalizations*. CRC Press.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (pp. 1322-1328). IEEE.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Ledoit, O., & Wolf, M. (2022). The power (non) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Multivariate Analysis*, *188*, Article 104844. <https://doi.org/10.1016/j.jmva.2021.104844>

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.

Lu, Y., & Yin, Y. (2022). Applying logistic lasso regression for the diagnosis of atypical Crohn's disease. *Computers in Biology and Medicine*, *141*, Article 105151. <https://doi.org/10.1016/j.combiomed.2021.105151>

Madera, J. (2008). Algoritmos evolutivos con estimación de distribuciones basados en pruebas de independencia.

Murphy, K. P. (2022). Probabilistic machine learning: An introduction. MIT Press.

Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012a). SMOTE-RSB*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using rough set theory. *Knowledge and Information Systems*, *33*(2), 245-265.

Richardson Ibáñez, J. (2017). Algoritmos evolutivos estimadores de distribución celulares para problemas de optimización continuos.

Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: Applications and issues. *Journal of Basic & Applied Sciences*, *13*, 459-465.

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

Zhang, C., Zhang, Y., & Zhang, Y. (2021). A survey on artificial intelligence for medical diagnosis. *Artificial Intelligence Review*, *54*(5), 3597-3645.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

Copyright © 2026, Autores: Leguen de Varona, Ireimis de las Mercedes, Madera Quintana, Julio César, Simón Cuevas, Alfredo, Rodríguez Guerra, Marcos Antonio



Esta obra está bajo una licencia de Creative Commons Atribución-No Comercial 4.0 Internacional