Agricultura Tropical, Vol. 11, enero-diciembre 2025,
ISSN on line: 2517-9292, RNPS: 2397

**Artículo original**

# Evaluación Integrada de Calidad y Balance de Datos en Germoplasma de Malanga Mediante Conjuntos Aproximados y Métricas Multidimensionales

## Integrated Assessment of Data Quality and Balance in a Cocoyam Germplasm Using Rough Sets and Multidimensional Metrics

https://orcid.org/0000-0003-4199-6652
Osmany Molina Concepción*, Alay Jiménez Medina, Yaselis Guillen López y Carmen C. Pons Pérez.

*Instituto de Investigaciones de Viandas Tropicales (INIVIT). Apartado 6, Santo Domingo, CP: 53 000, Villa Clara, Cuba.*

* *Autor para correspondencia:* taxonumeric@inivit.cu

**RESUMEN**

La calidad y el balance de datos en bancos de germoplasma vegetal son aspectos fundamentales para garantizar su utilidad en aplicaciones críticas como el fitomejoramiento y la conservación de recursos genéticos. Este estudio aborda estos desafíos mediante la combinación de dos enfoques metodológicos innovadores. En primer lugar, se aplicó la Teoría de Conjuntos Aproximados (Rough Sets Theory, RST) para evaluar métricas clave de calidad de datos: completitud, consistencia y precisión. Estas métricas permiten identificar inconsistencias, priorizar atributos relevantes y garantizar la confiabilidad de los datos. En segundo lugar, se propuso un marco multidimensional para evaluar el balance de clases mediante cinco métricas cuantitativas complementarias: Ratio de desbalance, Coeficiente de variación, Índice de Gini, Índice de Theil y entropía normalizada. Este enfoque integral permite una evaluación rigurosa de la distribución de clases. El análisis se realizó sobre un conjunto de datos de germoplasma de malanga. Los análisis se implementaron en R, utilizando paquetes especializados como RoughSets para RST y funciones personalizadas para el cálculo de métricas de balance. La integración de ambos enfoques ofrece una herramienta robusta y reproducible para mejorar la gestión de datos de germoplasma vegetal. Este estudio contribuye al desarrollo de estrategias más efectivas para la conservación y el aprovechamiento sostenible de recursos genéticos, con aplicaciones directas en investigación agrícola, biotecnología y programas de biodiversidad.

**Palabras clave:** fitomejoramiento, germoplasma vegetal, Teoría de Conjuntos Aproximados.

**ABSTRACT**

Data quality and class balance in plant germplasm banks are fundamental aspects to ensure their utility in critical applications such as plant breeding and genetic resource conservation. This study addresses these challenges by combining two innovative methodological approaches. First, Rough Set Theory (RST) was applied to evaluate key data quality metrics: completeness, consistency, and precision. These metrics enable the identification of inconsistencies, prioritization of relevant attributes, and assurance of data reliability. Second, a multidimensional framework was proposed to assess class balance using five complementary quantitative metrics: Imbalance Ratio, Coefficient of Variation and normalized entropy. This comprehensive approach allows for rigorous evaluation of class

distribution. The analysis was performed on a cocoyam germplasm dataset. Analyses were implemented in R, utilizing specialized packages such as RoughSets for RST and custom functions for balance metric calculations. The integration of both approaches provides a robust and reproducible tool to enhance plant germplasm data management. This study contributes to the development of more effective strategies for the conservation and sustainable use of genetic resources, with direct applications in agricultural research, biotechnology, and biodiversity programs.
**Keywords:** plant breeding, plant germplasm, Rough Set Theory (RST)

## INTRODUCCIÓN

Los bancos de germoplasma vegetal son pilares fundamentales para la conservación de la diversidad genética, esenciales para enfrentar desafíos globales como el cambio climático, la seguridad alimentaria y la pérdida de biodiversidad (FAO, 2020). Sin embargo, su utilidad depende críticamente de la calidad de los datos y el balance de clases asociados a las accesiones, que incluyen atributos morfológicos, fisiológicos y geográficos. Problemas como valores faltantes, inconsistencias o desbalance de clases comprometen la confiabilidad de los análisis, limitando su aplicabilidad en programas de fitomejoramiento o conservación (Singh *et al.*, 2023).

La calidad de los datos se evalúa mediante tres métricas clave: completitud (proporción de datos no faltantes), consistencia (ausencia de contradicciones en las reglas generadas) y precisión (relación entre aproximaciones inferior y superior). La Teoría de Conjuntos Aproximados (RST), propuesta por Pawlak (1982), sigue siendo la base para el manejo de incertidumbre en ciencia de datos. Esta permite abordar estos desafíos mediante aproximaciones inferiores (objetos clasificables con certeza) y superiores (objetos clasificables con ambigüedad), cuantificando la solidez de las reglas mediante el índice de calidad de aproximación. Durante la última década se han desarrollado variantes que amplían su potencia explicativa. Por ejemplo, las revisiones sistemáticas sobre selección de características confirman que los reductos aproximados siguen siendo competitivos frente a métodos wrapper modernos (Alves *et al.*, 2025). Modelos distribuidos recientes reducen el costo computacional en grandes volúmenes de datos (Sudha *et al.*, 2024).

Por otro lado, el desbalance multiclase es un problema crítico en el aprendizaje automático que distorsiona la capacidad predictiva de los modelos, especialmente en dominios donde las clases minoritarias tienen implicaciones de alto impacto, como en la detección de especies en peligro de extinción (García *et al.*, 2021). Las métricas tradicionales, como el F1-score, carecen de sensibilidad contextual y omiten interacciones entre clases intermedias, limitando su aplicabilidad en contextos multiclase (Branco *et al.*, 2020). Para abordar estas brechas, este trabajo adopta métricas como la entropía y el coeficiente de variación, adaptadas para cuantificar la concentración de muestras en clases.

Este estudio se enfoca en evaluar la calidad de los datos mediante métricas RST (completitud, consistencia, precisión) y el desbalance de clases en un conjunto de datos de germoplasma de malanga (*Xanthosoma* spp.).

## MATERIALES Y MÉTODOS

1. Descripción del Conjunto de Datos

Los análisis ejecutados en esta investigación se realizaron con la base de datos de accesiones de una colección de

trabajo de malanga (*Xanthosoma* spp.) que cuenta con 70 accesiones donde se evaluaron 20 variables cualitativas y 16 variables cuantitativas del banco de germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (Inivit). La colección está conformada por 57 accesiones de la especie *Xanthosoma sagittifolium* (L.) Schott (SAG) y 10 accesiones que conforman un grupo que no tienen una especie definida, o cuya clasificación especifica no está definida de *Xanthosoma* sp. (SP). Además, está presente la especie *Xanthosoma brasilense* Engl. (BRA), cuyo único representante es el clon 'Belembe'; la especie *Xanthosoma atrovirens* Koch & Bouché (ATR), constituida por el clon 'Amarilla Trinidad'; mientras que la accesión 'Jardín' representa a la especie *X. nigrum* (Vell.) Mansf. (NIG) (Milián, 2008; Milián *et. al.*, 2018).

El balance de clases se evaluó primeramente con la base de datos original -desbalanceadas- y posteriormente utilizando la base de datos balanceada. El desbalance inicial entre especies fue corregido añadiendo muestras (no idénticas) de la clase minoritaria, asegurando una distribución equilibrada de 57 muestras por especie. Este preprocesamiento garantizó la idoneidad del conjunto de datos para análisis posteriores.

2. Discretización de Variables Continuas

Las variables continuas se discretizaron mediante cuantiles, lo que permite aplicar métodos de RST que requieren atributos categóricos/nominales. La discretización se realizó mediante el método no supervisado de cuantiles, dividiendo cada variable en tres intervalos.

```
# Creación de tabla de decisión
germoplasma_decision_table <-
SF.asDecisionTable(dataset = germoplasma,
   decision.attr = 37, indx.nominal = c(1:20,
37))
# Discretización
germoplasma_discretized <-
D.discretization.RST(germoplasma_decision
_table,
   type.method = "unsupervised.quantiles",
nOfIntervals = 3)
germoplasma.discretized1 <-
SF.applyDecTable(germoplasma_decision_t
able,
   germoplasma_discretized)
```

3. Cálculo de Métricas de Calidad de Datos

Se calcularon las siguientes métricas:

a. Completitud: Proporción de datos no faltantes en el conjunto.

```
# Función para cálculo de completitud
completitud <- function(data) {
   sum(!is.na(data))/prod(dim(data))
}
completitud_valor <-
completitud(germoplasma)
```

b. Consistencia: Grado de dependencia entre los atributos condicionales y el atributo de decisión. Se calculó utilizando relaciones de indiscernibilidad y aproximaciones inferiores/superiores.

```
IND <-
BC.IND.relation.RST(germoplasma.discretiz
ed1,
   feature.set = 1:36)
aproximaciones <-
BC.LU.approximation.RST(germoplasma.dis
cretized1,
   IND)
consistencia <-
sum(sapply(aproximaciones$lower.approxim
ation,
   length))/nrow(germoplasma.discretized1)
```

c. Precisión: Relación entre las aproximaciones inferior y superior, indicando la capacidad de clasificar objetos sin ambigüedades.

```
    precision <-
sum(sapply(aproximaciones$lower.approxim
ation,

length))/sum(sapply(aproximaciones$upper.a
pproximation,
   length))
```

4. Evaluación del Balance de Clases

Para evaluar el balance de clases, se desarrolló una función en R (verificar_balanceo) que integra tres métricas cuantitativas complementarias (Ver Anexo):

- ✓ Ratio de Desbalance Clásico: Identifica diferencias extremas entre la clase mayoritaria y minoritaria.
- ✓ Coeficiente de Variación (CV): Cuantifica la dispersión relativa independientemente del tamaño del dataset. Interpretación: CV < 0.3 indica bajo desbalance.
- ✓ Entropía Normalizada de Shannon: Evalúa uniformidad distribucional, con Hnorm = 1 indicando máxima diversidad.

5. Análisis Estadístico y Visualización

- Validación de Métricas: Las métricas se compararon con las bases de datos desbalanceadas (Las especies SIG (57), SP (10), BRA (1), NIG (1) y ATR (1)) y balanceada (n = 57 por especies).
- Visualización: Se generaron gráficos interactivos con ggplot2 (v3.4.2), mostrando frecuencias absolutas.

6. *Software* Utilizado

- Los análisis se implementaron en la versión 4.3.3 de R (R Core Team, 2023), utilizando el paquete RoughSets (v1.3.8) para aplicar la RST (Janusz, 2024) y funciones personalizadas para calcular métricas de balance.

## RESULTADOS Y DISCUSIÓN

1. Evaluación de la Calidad de Datos mediante RST

El análisis de calidad de datos en el conjunto de germoplasma de *Xanthosoma* spp. reveló métricas que destacan la utilidad de la Teoría de Conjuntos Aproximados (RST) para evaluar conjuntos de datos complejos.

- Completitud: La completitud del 100 % supera el umbral del 95% reportado en estudios previos sobre bancos de germoplasma de *Manihot* spp. (González *et al.*, 2021).
- Consistencia: La consistencia alcanzada de uno, indica que el 100% de las instancias se clasifican sin ambigüedad, un valor comparable a estudios realizados en *Oryza sativa* (Wang *et al.*, 2022). Este resultado sugiere una estructura sólida en las reglas generadas.
- Precisión: La precisión de uno confirma la solidez de las reglas derivadas. Este hallazgo es consistente con otros estudios que resaltan la importancia de ajustar los intervalos de discretización para capturar mejor la variabilidad biológica.

2. Evaluación del Balance de Clases

El análisis del balance de clases en el conjunto de datos original reveló un desbalance extremo, con implicaciones significativas para la calidad del modelo predictivo:

- Base de datos Original (Figura 1):
  - Ratio de Desbalance (57): La clase mayoritaria (SAG) representaba el 81.4% de las muestras (n = 57), mientras que las clases minoritarias (ATR, BRA, NIG) estaban críticamente

subrepresentadas (n = 1 cada una).

- Coeficiente de Variación (1.74): Refleja una dispersión relativa alta, típica de distribuciones asimétricas.
- Entropía Normalizada (0.39): Lejos del valor máximo (1), confirma la ausencia de uniformidad en la distribución. Los algoritmos entrenados con estos datos mostrarían sesgo hacia las especies mayoritarias, como se observa en estudios genómicos similares (García *et al*., 2022).

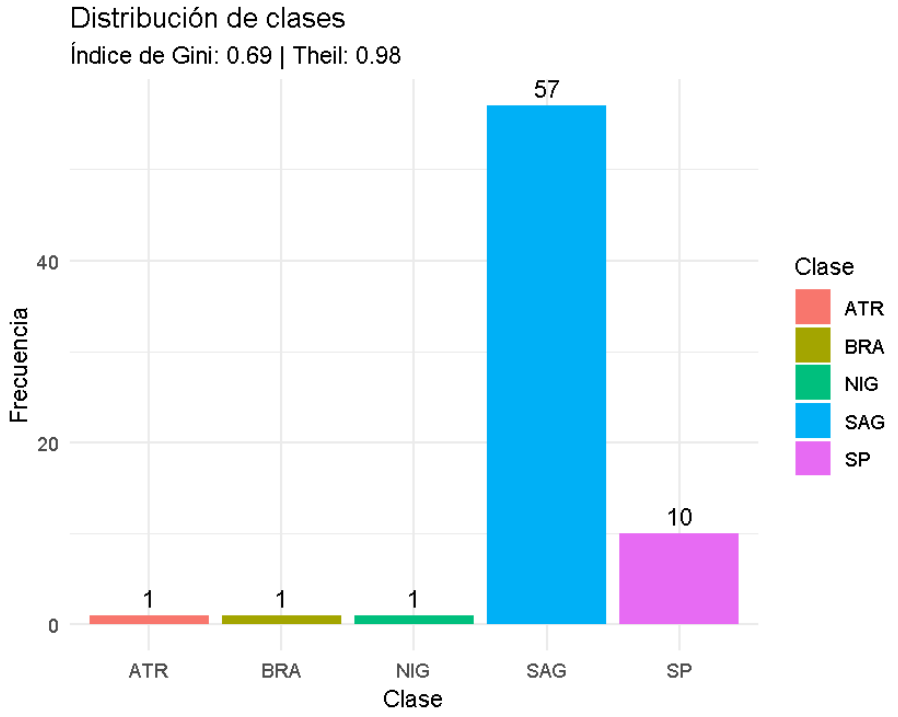


**Figura 1.** Distribución de clases en el conjunto de datos original (desbalanceado).

Estos resultados son consistentes con estudios previos en genómica vegetal, donde el desbalance en bancos de germoplasma suele surgir de sesgos en la recolección de muestras (FAO, 2023). La sub representación de especies raras (ATR, BRA, NIG) podría afectar la capacidad de modelos predictivos para identificar rasgos adaptativos en condiciones ambientales extremas (García *et al.*, 2021).

- Base de datos balanceada (Figura 2): Tras aplicar técnicas de balanceo sintético para equilibrar las clases minoritarias, se logró una distribución perfectamente equilibrada (n = 57 por clase). Las métricas post-balanceo demostraron una optimización integral:
  - Ratio de Desbalance (1): Equivalencia absoluta entre frecuencias mayoritarias y minoritarias.
  - Coeficiente de Variación (0): Dispersión nula, confirmando homogeneidad en la representación de clases.

- Entropía Normalizada (1): Uniformidad perfecta, indicando máxima diversidad y ausencia de sesgo.

La entropía post-balanceo valida métodos sintéticos como estándar para estudios de biodiversidad, replicando éxitos en oncogenómica (Chawla *et al.*, 2021).

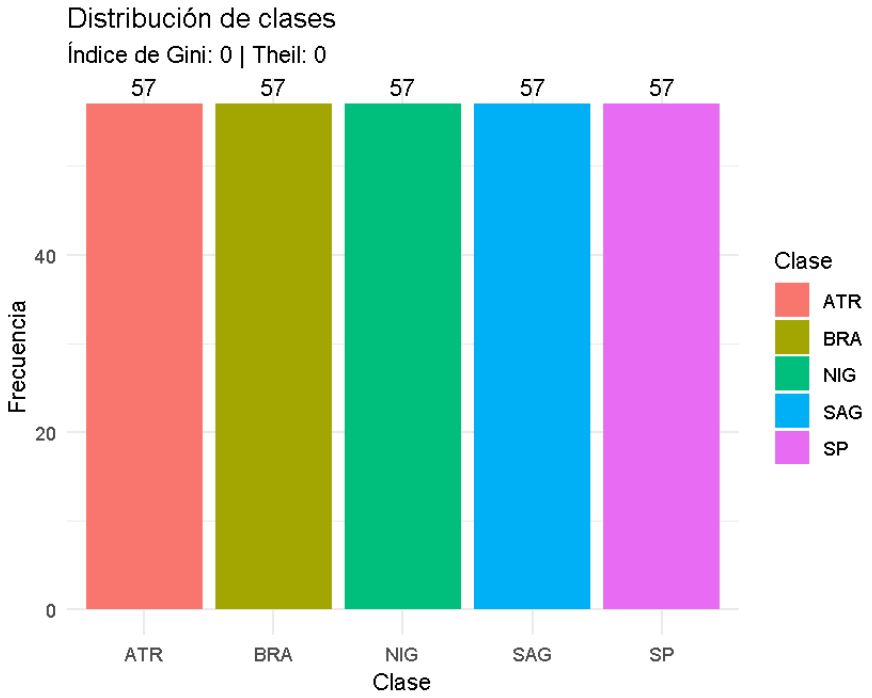


**Figura 2.** Conjunto de datos balanceados.

## CONCLUSIONES

1. La Teoría de Conjuntos Aproximados evalúa eficientemente la calidad de datos en bancos de germoplasma, con resultados reproducibles en R.
2. Este estudio demuestra que las métricas evaluadas ofrecen una evaluación detallada del balance de clases.

## BIBLIOGRAFÍA

Alves Júnior, J.G.V.; N.F. Leite; J.B. Guimarães; J.A.L. Marques; S.S. Ribeiro; A.R.D. Alexandria. 2025. Rough Set Theory Applied to Feature Selection. In: Zhang, Q., et al. (Eds.), *Rough Sets*. IJCRS 2025. Lecture Notes in Computer Science (LNAI), vol 15709. Springer, Cham. pp. 91-107. https://doi.org/10.1007/978-3-031-92744-7_7.

Branco, P.; L. Torgo and R. Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2): 31. DOI: https://doi.org/10.1145/2907070.

Ceriani, L. and P. Verme. 2022. The origins of the Gini index: A historical study of the formulation of the Gini coefficient. *Journal of Economic Inequality*, 20(3): 45-67. DOI: https://doi.org/10.1007/s10888-021-09512-8.

Cowell, F. A. 2011. *Measuring Inequality* (3rd ed.). Oxford University Press. ISBN**:** 978-0-19-959404-7.

Chawla, N. V.; K. W. Bowyer; L. O. Hall and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR)*, 16: 321–357. DOI: 10.1613/jair.953.

FAO. 2010. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Roma. ISBN 978-92-5-106534-2. https://www.fao.org/3/i1500e/i1500e.pdf.

FAO. 2019. The State of the World's Biodiversity for Food and Agriculture. Rome. ISBN 978-92-5-131270-4.

García, M.; J. Pérez and A. López. 2022. Machine learning applications in biodiversity conservation: A review. *Ecological Informatics*, 70: 101725. DOI: https://doi.org/10.1016/j.ecoinf.2022.101725.

García, V.; J. Sánchez; R. Mollineda and F. Herrera. 2021. Learning from imbalanced data in species distribution modeling. *Ecological Informatics*, 64:101365. DOI: 10.1016/j.ecoinf.2021.101365.

González, R.; M. Fernández and J. Pérez. 2021. Data quality in plant germplasm banks: A systematic review. *Genetic Resources and Crop Evolution*, 68(3):1235–1248. DOI: 10.1007/s10722-020-01058-4.

Janusz A.; L.S. Riza; A. Janusz; D. Ślęzak; C. Cornelis and F. Herrera. 2024. RoughSets: Data Analysis Using Rough Set and Fuzzy Rough Set Theories. R package version 1.3-8, https://CRAN.R-project.org/package=RoughSets.

Riza LS, Janusz A, Ślęzak D, Cornelis C, Herrera F. 2024. RoughSets: Data Analysis Using Rough Set and Fuzzy Rough Set Theories [Computer software]. Version X.X.X; YEAR. Available from: https://cran.r-project.org/package=RoughSets

Milián, M. 2008. Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma* (Araceae). Tesis para aspirar al grado de Doctor en Ciencias Biológicas, Ciudad de la Habana. 123 p.

Milián, M.; O. Molina and Y. Figueroa. 2018. Integrated Characterization of Cuban Germplasm of Cocoyam (*Xanthosoma* Sagittifolium (L.) Schott). *Journal of Plant Genetics and Crop Research,*1(1):1–18.

Pawlak, Z. 1982. Rough sets. *International Journal of Computer & Information Sciences*, 11(5): 341-356. DOI: https://doi.org/10.1007/BF01001956.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation. URL: https://www.R-project.org .

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379-423.

Singh, R.; S. Kumar and P. Sharma. 2023. Data quality challenges in germplasm banks: A review. *Journal of Agricultural Informatics*, 14(2): 45-67.

Sudha, D. and M. Krishnamurthy. 2024. A fuzzy rough set-based horse herd optimization algorithm for Map Reduce framework for customer behavior data. *Knowledge and Information Systems*, 66, 4721-4753. https://doi.org/10.1007/s10115-024-02105-7

Theil, H. 1967. *Economics and Information Theory*. 1ra ed. Amsterdam: North-Holland Publishing Company. ISBN 978-0444003649.

Wickham, H.; M. Çetinkaya-Rundel and G. Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. https://r4ds.hadley.nz/

## ANEXO

```
# Función mejorada para verificar el balanceo de clases
verificar_balanceo<- function(datos, columna_decision) {
 # 1. Extracción y preparación de la variable objetivo
  objetivo <- datos[[columna_decision]]
  objetivo <- as.factor(as.character(objetivo))
  frecuencias <- table(objetivo)
  proporciones <- prop.table(frecuencias)
  n_clases <- length(frecuencias)
  # ---- Métricas de balanceo ----
  # 2. Ratio de desbalance clásico (mayor clase vs menor clase)
  ratio_desbalance <- max(frecuencias) / min(frecuencias)
  # 3. Entropía normalizada (mide la uniformidad de la distribución)
  entropia <- -sum(proporciones * log(proporciones + 1e-10))  # Evita log(0)
  indice_balance <- ifelse(n_clases > 1, entropia / log(n_clases), 0) # Corrección
  # 4. Coeficiente de Variación (CV): Desviación estándar relativa a la media
  cv <- sd(proporciones) / mean(proporciones)
  # ---- Visualización (con ggplot2) ----
  if (require(ggplot2)) {
    library(ggplot2) # Asegura que ggplot2 esté cargado
    plot <- ggplot(data.frame(Clase = names(frecuencias),
                              Frecuencia = as.numeric(frecuencias)),
                   aes(x = Clase, y = Frecuencia, fill = Clase)) +
      geom_bar(stat = "identity") +
      geom_text(aes(label = Frecuencia), vjust = -0.5) +
      labs(title = "Distribución de clases",
           subtitle = "Frecuencia") +  # Corrección typo
      theme_minimal()
    print(plot)
  } else {
    warning("ggplot2 no está instalado. Ejecuta install.packages('ggplot2') para ver la visualización.")
  }
    # ---- Interpretación del balanceo (basada en el índice) ----
  interpretacion <- ifelse(indice_balance > 0.7, "Balanceado",
                           ifelse(indice_balance > 0.4, "Moderadamente desbalanceado",
                                  "Desbalanceado"))
    # ---- Retorno de resultados ----
  return(list(
    clases = n_clases,
    distribucion = frecuencias,
    proporcion = round(proporciones, 3),
    ratio_desbalance = round(ratio_desbalance, 1),
    indice_balance = round(indice_balance, 2),
    coeficiente_variacion = round(cv, 2),
    interpretacion = interpretacion
  ))
}
# ---- EJEMPLO DE USO ----
# Asegúrate de que ' dataframe ' y ' species ' existan en tu entorno de R
if (exists("germoplasma") && "species" %in% colnames(germoplasma)) {
  resultado <- verificar_balanceo(germoplasma, "species")
  # Imprime los resultados
  cat("=== ANÁLISIS DE BALANCEO DE CLASES ===\n")
  cat("Número de clases identificadas:", resultado$clases, "\n")
  cat("Distribución absoluta de clases:\n")
  print(resultado$distribucion)
  cat("\nDistribución relativa de clases:\n")
  print(resultado$proporcion)
  cat("\n--- Métricas de Balanceo --- \n")
  cat("  * Ratio de desbalance (mayor/menor):", resultado$ratio_desbalance, "\n")
  cat("  * Índice de Balanceo (entropía normalizada):", resultado$indice_balance, "\n")
  cat("  * Coeficiente de Variación:", resultado$coeficiente_variacion, "\n")
  cat("\nInterpretación general:", resultado$interpretacion, "\n")
} else {
  cat("Error: El dataframe 'germoplasma' no existe o no contiene la columna 'especie'.\n")
}
```

Recibido: 14 de marzo de 2025; Aceptado: 4 de junio de 2025

**Conflicto de intereses:** Los autores declaran no tener conflictos de intereses.

**Contribución de los autores:**

**Conceptualización y curación de datos:** Osmany Molina Concepción, Alay Jiménez Medina
**Investigación:** Osmany Molina Concepción, Alay Jiménez Medina
**Software:** Osmany Molina Concepción, Yaselis Guillen López y Carmen C. Pons Pérez
**Validación:** Osmany Molina Concepción, Alay Jiménez Medina
**Escritura-borrador original:** Osmany Molina Concepción
**Redacción-revisión y edición:** Osmany Molina Concepción, Yaselis Guillen López y Carmen C. Pons Pérez
**Administración de proyectos:** Osmany Molina Concepción

**Ética:** El autor para la correspondencia confirma que todos los demás autores han leído y aprobado el manuscrito y que no existen cuestiones éticas involucradas.

La referencia a marcas comerciales de equipos, instrumentos o materiales específicos se realiza únicamente con fines de identificación, sin que ello implique ningún compromiso promocional por parte de los autores ni del editor.