Evaluación de métricas en modelos predictivos de clasificación en Machine Learning

John Jairo Campo Yepes

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

27/ene/2026

Evaluación de métricas en modelos predictivos de clasificación en Machine Learning

John Jairo Campo Yepes

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ciencia de Datos

Director (a):

Diego Armando García García

Modalidad:

Monografía

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

27/ene/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 27/ene/2026

A la Universidad Ean por brindarme la oportunidad de continuar formando profesionalmente como Magister.

La locura es hacer la misma cosa una y otra vez esperando obtener resultados diferentes.
Albert Einstein.

Agradecimientos

Expreso mi más sincero agradecimiento a la Universidad EAN por brindarme la oportunidad de formarme académicamente y alcanzar el título de Magíster, así como por el acompañamiento institucional y académico recibido a lo largo de este proceso de formación, el cual contribuyó de manera significativa a mi crecimiento profesional y personal.

Expreso un sincero agradecimiento a mi tutor de trabajo de grado, Diego Armando García, por su acompañamiento, orientación académica y valiosas observaciones a lo largo del desarrollo de esta investigación. Su experiencia y rigor metodológico fueron fundamentales para fortalecer la calidad y el enfoque del presente trabajo.

De manera especial, agradezco a mi esposa Heydi Gómez, a mis hijas Angeline Campo y Nicolle Campo, y a mi madre Rosario Yepes, quienes con su apoyo incondicional, comprensión y constante motivación hicieron posible la culminación de este proyecto académico. Su paciencia, aliento y confianza fueron fundamentales para superar los retos que implicó este proceso, y constituyen el principal sustento de este logro. Este trabajo es, en gran medida, resultado de su respaldo y acompañamiento permanente.

Resumen

El propósito de este estudio fue analizar el comportamiento y la estabilidad de diversas métricas de evaluación utilizadas en modelos de clasificación, considerando su sensibilidad frente a diferentes niveles de desbalance en los datos. Para ello, se definieron tres escenarios de diseño experimental computacional: uno balanceado, uno moderadamente desbalanceado y otro con un desbalance extremo. En cada caso se entrenaron modelos de aprendizaje automático y se calcularon métricas tradicionales como Accuracy, Precision, Recall, F1-score y AUC, con el fin de identificar variaciones en su desempeño. Adicionalmente, se aplicaron pruebas estadísticas no paramétricas, específicamente el test de rangos con signo de Wilcoxon, para comparar las métricas sin asumir supuestos de normalidad.

Los resultados evidencian que, en escenarios balanceados, la mayoría de las métricas presentan comportamientos estables y diferencias reducidas. No obstante, a medida que aumenta el desbalance, algunas métricas pierden confiabilidad, en particular Accuracy y Precision, que tienden a sobreestimar el rendimiento del modelo. En contraste, AUC mostró mayor consistencia a lo largo de los escenarios analizados, mientras que Recall y F1-score reflejaron una mayor sensibilidad a la baja prevalencia de la clase minoritaria. El análisis estadístico permitió identificar diferencias significativas entre métricas en escenarios desbalanceados, lo que respalda las hipótesis planteadas y es desarrollado en detalle a lo largo del documento.

En conclusión, el estudio resalta la importancia de seleccionar métricas de evaluación acordes con las características del conjunto de datos y destaca la utilidad de las pruebas no paramétricas como herramienta robusta para la comparación de métricas en problemas de clasificación.

Palabras clave:

- Métricas de evaluación
- Desequilibrio de clases
- Aprendizaje automático
- Pruebas no paramétricas
- Rendimiento del modelo
- Clasificación supervisada
- Validación estadística

Abstract

The purpose of this study was to analyze the behavior and stability of several evaluation metrics used in classification models, considering their sensitivity to different levels of class imbalance. To this end, three computational experimental scenarios were defined: a balanced scenario, a moderately imbalanced scenario, and a highly imbalanced scenario. In each case, machine learning classification models were trained, and traditional performance metrics such as Accuracy, Precision, Recall, F1-score, and AUC were computed in order to examine variations in their behavior. Additionally, non-parametric statistical tests, specifically the Wilcoxon signed-rank test, were applied to compare metrics without assuming normality.

The results indicate that, under balanced conditions, most metrics exhibit stable behavior with minor differences among them. However, as class imbalance increases, certain metrics particularly Accuracy and Precision become less reliable and tend to overestimate model performance. In contrast, AUC showed greater consistency across the analyzed scenarios, while Recall and F1-score demonstrated higher sensitivity to the low prevalence of the minority class. The statistical analysis revealed significant differences among metrics in imbalanced scenarios, supporting the hypotheses proposed and discussed in detail throughout the study.

In conclusion, this research highlights the importance of selecting evaluation metrics according to the characteristics of the dataset and underscores the usefulness of non-parametric statistical tests as a robust approach for comparing classification metrics.

Keywords:

- Evaluation metrics
- Class imbalance
- Machine learning

- Nonparametric tests
- Model performance
- Supervised classification
- Statistical validation

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Introducción

El aprendizaje automático (Machine Learning, ML) se ha consolidado como un componente esencial en la transformación digital de múltiples sectores, al permitir la construcción de modelos predictivos capaces de aprender patrones complejos a partir de grandes volúmenes de datos (Jordan & Mitchell, 2015). Su aplicación en áreas como la salud, las finanzas, la industria y los servicios públicos ha generado avances significativos en la automatización de procesos, la toma de decisiones basada en datos y la optimización operativa. Sin embargo, este crecimiento ha evidenciado la necesidad de evaluar rigurosamente el rendimiento de los modelos, puesto que su comportamiento puede variar de manera considerable dependiendo del tipo de datos, el algoritmo utilizado y las condiciones del entorno en el que operan.

La evaluación de modelos predictivos se realiza mediante métricas que cuantifican su desempeño. No obstante, la selección de dichas métricas no es trivial. En la práctica, se observa una tendencia a utilizar indicadores tradicionales como la exactitud (accuracy), incluso en escenarios donde resultan insuficientes o generan interpretaciones sesgadas (Saito & Rehmsmeier, 2015; Chicco & Jurman, 2020). Este problema es especialmente crítico en contextos con clases desbalanceadas, alta variabilidad en los errores o restricciones éticas, donde métricas como F1-score o AUC ofrecen perspectivas más adecuadas (Chai & Draxler, 2014; Mehrabi et al., 2021).

Adicionalmente, el auge de enfoques orientados a la interpretabilidad (SHAP, LIME) y la equidad algorítmica ha introducido nuevas exigencias para la evaluación de modelos, dado que estas dimensiones pueden modificar la percepción del desempeño y la aceptabilidad del sistema (Molnar, 2022; Verma & Rubin, 2018). Estas consideraciones evidencian la necesidad de contar con un análisis comparativo y contextualizado que permita identificar la pertinencia, robustez y coherencia de las métricas en distintos escenarios.

A partir de esta problemática, surge la necesidad de comprender cuáles métricas resultan más adecuadas para evaluar el desempeño de modelos predictivos de clasificación, considerando las características de los datos, el tipo de problema y el contexto de aplicación. De este modo, la presente investigación se orienta por la siguiente pregunta:

¿Qué métricas permiten evaluar de manera más adecuada el desempeño de modelos predictivos de clasificación, en función del tipo de problema, los datos y el contexto de aplicación?

Para dar respuesta a esta pregunta, se propone un análisis comparativo sustentado en pruebas estadísticas no paramétricas como el Wilcoxon Signed-Rank test que permiten identificar diferencias significativas entre métricas sin depender de supuestos de normalidad. Este enfoque busca aportar una guía estructurada para la selección de métricas, integrando dimensiones técnicas, operativas y éticas.

El documento se organiza en ocho capítulos. El primero presenta los objetivos y la justificación de la investigación. El segundo desarrolla el marco teórico relacionado con el aprendizaje automático, los modelos predictivos y las métricas de evaluación. El tercer capítulo expone las hipótesis y variables del estudio. El cuarto describe la metodología, incluyendo el diseño, la población, los instrumentos y las técnicas de análisis. El quinto detalla el trabajo de campo y el procesamiento de datos, mientras que el sexto presenta los resultados obtenidos. El séptimo capítulo expone la discusión y las implicaciones del estudio. Finalmente, el octavo capítulo describe las conclusiones y las líneas de trabajo futuro.

Objetivos

Objetivo general

Analizar y comparar el desempeño de diversas métricas empleadas en modelos de clasificación mediante pruebas estadísticas no paramétricas, con el propósito de determinar su consistencia, robustez y pertinencia en diferentes contextos y tipos de datos.

Objetivos específicos

1. Caracterizar las métricas de evaluación utilizadas en modelos de clasificación, considerando sus fundamentos conceptuales, propiedades y limitaciones.
2. Analizar el comportamiento de métricas de evaluación en problemas de clasificación supervisada, utilizando modelos de referencia, bajo distintos escenarios de desbalance de clases.
3. Aplicar y comparar diversas métricas de rendimiento sobre los modelos entrenados, examinando su comportamiento frente a distintos niveles de balanceo, complejidad y variabilidad.
4. Evaluar estadísticamente las diferencias entre métricas mediante pruebas no paramétricas, identificando patrones de superioridad, equivalencia o inconsistencia.
5. Interpretar y sintetizar los resultados obtenidos para formular recomendaciones fundamentadas sobre la selección de métricas en función del tipo de problema, el contexto de aplicación y los requisitos analíticos del modelo.

Justificación

La evaluación del rendimiento en los modelos de aprendizaje automático constituye un componente esencial para garantizar la validez, confiabilidad y aplicabilidad de las soluciones basadas en datos. La elección adecuada de métricas no solo determina la calidad de las predicciones, sino que incide directamente en la interpretación de los resultados, la toma de decisiones y la implementación de los modelos en escenarios reales. En este contexto, la literatura ha señalado que la evaluación del desempeño de modelos de clasificación resulta particularmente desafiante cuando la distribución de clases es desbalanceada, ya que métricas tradicionales como la accuracy pueden conducir a interpretaciones sesgadas al privilegiar la clase mayoritaria y subestimar el comportamiento del modelo sobre la clase minoritaria (Japkowicz, 2000; He & Garcia, 2009).

A pesar de estas limitaciones ampliamente documentadas, en la práctica persiste la tendencia a utilizar métricas tradicionales (como accuracy) incluso en situaciones donde resultan insuficientes, especialmente en contextos de desbalance de clases, alta variabilidad de los errores o problemas con restricciones operativas y éticas. Esta situación evidencia una brecha entre las necesidades actuales de la disciplina y los criterios empleados en la selección de métricas de evaluación. En respuesta a esta problemática, diversos estudios han propuesto el uso de métricas alternativas como precision, recall, F1-score y AUC, las cuales permiten una evaluación más informativa y contextualizada del rendimiento del clasificador en escenarios de desbalance (Powers, 2011; Saito & Rehmsmeier, 2015). No obstante, la selección e interpretación de estas métricas no es trivial y depende del contexto del problema y de los costos asociados a los distintos tipos de error.

La presente investigación se justifica, por tanto, en la necesidad de establecer un análisis comparativo sistemático y metodológicamente riguroso que permita comprender el comportamiento y la pertinencia de distintas métricas de evaluación en tareas de clasificación. El uso de pruebas estadísticas no paramétricas aporta robustez al análisis, al permitir contrastar el rendimiento de las métricas sin depender de supuestos de normalidad y considerando su variabilidad bajo múltiples escenarios, tal como se recomienda en la literatura para la comparación de métodos de aprendizaje automático (Demšar, 2006). De este modo, el estudio contribuye a la formulación de criterios más precisos y fundamentados para la selección de métricas, aspecto clave tanto en la práctica profesional como en el ámbito académico.

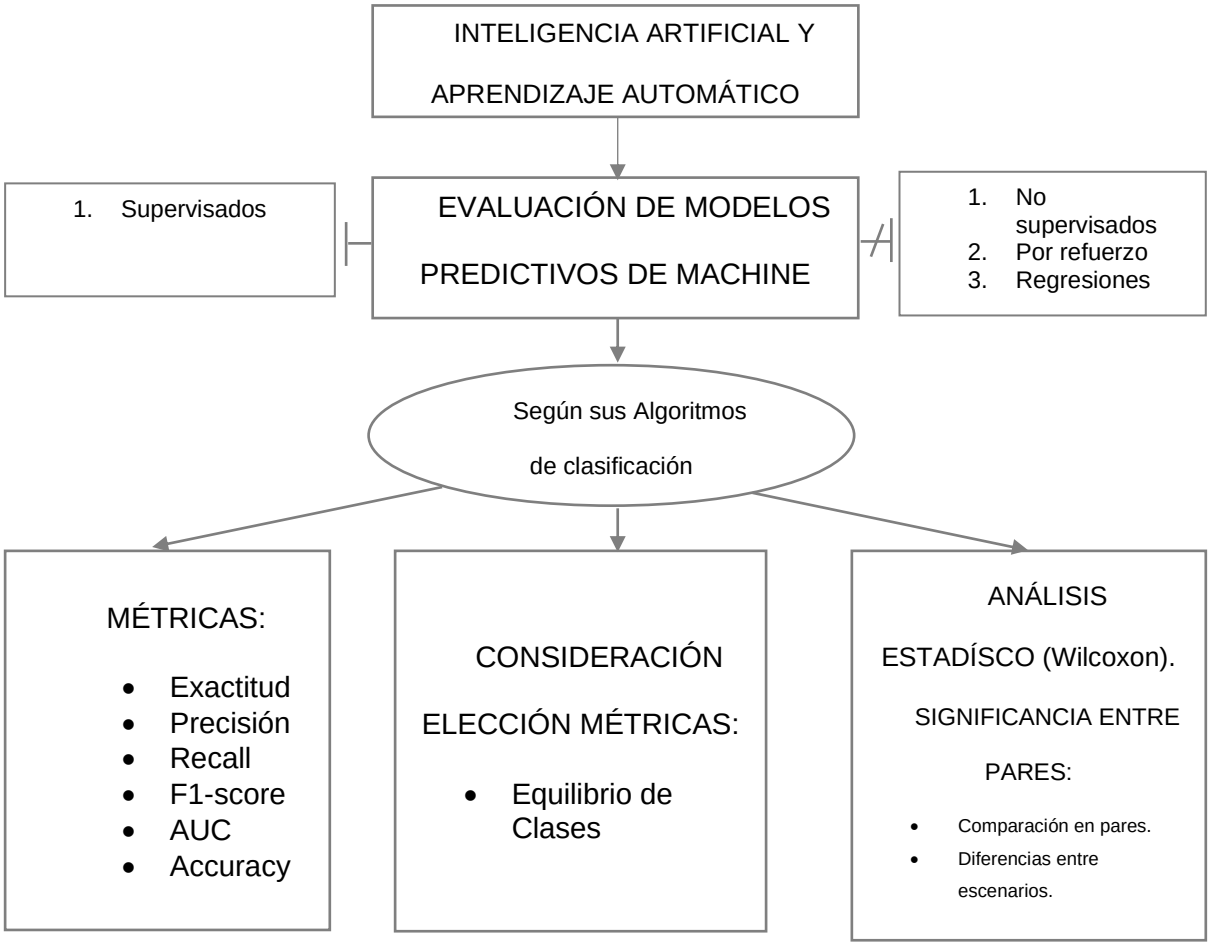
Adicionalmente, el creciente interés por la interpretabilidad, la equidad algorítmica y la transparencia en los sistemas de inteligencia artificial exige ampliar la evaluación más allá de un único valor numérico de desempeño. Este enfoque integral resulta particularmente relevante para organizaciones que buscan modelos responsables, auditables y alineados con principios éticos. En este sentido, la investigación aporta lineamientos que integran dimensiones técnicas y éticas, orientados al fortalecimiento del uso responsable del aprendizaje automático en diversos contextos institucionales, científicos y empresariales, contribuyendo así a la consolidación de prácticas más informadas, coherentes y contextualizadas en la evaluación de modelos predictivos.

Marco Teórico

El aprendizaje automático (Machine Learning, ML) constituye un conjunto de métodos y algoritmos diseñados para identificar patrones en los datos y realizar predicciones o decisiones sin una programación explícita para cada tarea específica. Su adopción se ha intensificado en diversos sectores, impulsada por la disponibilidad de grandes volúmenes de información, avances computacionales y la necesidad de automatizar procesos complejos. Dentro de este campo, las tareas de clasificación representan los problemas más relevantes, ya que permiten estimar categorías a partir de variables observadas.

Figura 1

Mapa conceptual evaluación de los modelos de clasificación



Nota. Elaboración propia

Aprendizaje automático y modelos predictivos

El ML se fundamenta en la construcción de modelos capaces de generalizar a nuevos datos mediante el aprendizaje a partir de ejemplos. Un modelo predictivo busca establecer una relación entre las características de entrada y el valor objetivo, minimizando el error entre las predicciones y los resultados reales. Para ello, existen múltiples enfoques algorítmicos, entre los que destacan los modelos lineales, árboles de decisión, máquinas de vectores de soporte, redes neuronales y métodos de ensamblaje. La elección del algoritmo depende de factores como la naturaleza de los datos, el nivel de complejidad requerido, la capacidad de interpretabilidad y las restricciones del contexto de aplicación, tal como se discute en la literatura clásica y contemporánea sobre aprendizaje automático (Mitchell, 1997; Bishop, 2006; Hastie, Tibshirani & Friedman, 2009).

Problemas de clasificación

En un problema de clasificación, el objetivo es asignar a cada instancia una etiqueta perteneciente a un conjunto predefinido de categorías. Estos modelos son utilizados en aplicaciones como detección de fraude, diagnóstico médico, sistemas de recomendación, análisis de sentimiento y filtrado de spam. La complejidad aumenta cuando existen clases desbalanceadas, es decir, cuando unas categorías están representadas en mayor proporción que otras, situación que tiende a sesgar los modelos hacia las clases mayoritarias (Japkowicz, 2000; He & Garcia, 2009).

Métricas de evaluación para clasificación

La evaluación del rendimiento en clasificación requiere métricas que cuantifiquen aspectos distintos del desempeño. Algunas de las más utilizadas incluyen:

- Exactitud (Accuracy): Proporción de predicciones correctas sobre el total. Aunque popular, puede resultar engañosa en casos de desbalance de clases.

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

- Precisión: Mide la proporción de verdaderos positivos entre las predicciones positivas, útil en escenarios donde los falsos positivos son costosos.

$$Precisión = \frac{VP}{VP + FP}$$

- Sensibilidad o Recall: Indica la capacidad del modelo para identificar correctamente los casos positivos, especialmente relevante cuando los falsos negativos representan un riesgo.

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

- Especificidad (Specificity): Evalúa la capacidad del modelo para identificar correctamente los casos negativos. Es útil cuando se desea minimizar los falsos positivos.

$$Especifidad = \frac{VN}{VN + FP}$$

- F1-score: Media armónica entre precisión y recall, recomendada cuando se requiere un equilibrio entre ambas métricas.

$$F1 = \frac{2 * Precisión * Sensibilidad}{Precisión + Sensibilidad}$$

- AUC-ROC: Evalúa la capacidad del modelo para discriminar entre clases al variar los umbrales de decisión, siendo menos sensible al desbalance.

Estas métricas capturan dimensiones diferentes del comportamiento del modelo, por lo que su interpretación debe alinearse con la naturaleza del problema y los costos asociados a los errores.

Las definiciones y descripciones de las métricas de desempeño utilizadas en este estudio se elaboraron con base en los trabajos de Saito y Rehmsmeier (2015) y Chicco y Jurman (2020), los cuales analizan el comportamiento de métricas de clasificación en contextos con diferentes niveles de desbalance de clases.

Estas métricas capturan dimensiones diferentes del comportamiento del modelo, por lo que su interpretación debe alinearse con la naturaleza del problema y los costos asociados a los errores. La literatura enfatiza que ninguna métrica es universalmente superior, sino que su pertinencia depende del contexto del problema y de la distribución de las clases (Powers, 2011; Chicco & Jurman, 2020).

El desbalance de clases y su impacto en las métricas de evaluación

El desbalance de clases es uno de los retos más frecuentes en la clasificación supervisada y ocurre cuando la distribución entre clases es altamente asimétrica. Este fenómeno afecta de manera directa el desempeño de los modelos y distorsiona la interpretación de varias métricas tradicionales (He & Garcia, 2009). En dominios como detección de fraude, diagnóstico médico o riesgo crediticio, la clase minoritaria suele representar un porcentaje muy reducido de los datos, lo que genera sesgos tanto en el entrenamiento como en la evaluación.

En primer lugar, Accuracy tiende a sobreestimar el rendimiento del modelo en escenarios desbalanceados. Un clasificador que predice siempre la clase mayoritaria puede alcanzar valores altos de exactitud sin capturar patrones relevantes (Japkowicz & Stephen, 2002). Por esta razón, Accuracy es considerada una métrica poco confiable cuando la prevalencia de la clase positiva es baja.

Por otro lado, Precision se ve fuertemente afectada por la cantidad de falsos positivos. Cuando la clase minoritaria es escasa, cualquier predicción incorrecta hacia la clase positiva produce una caída significativa en el puntaje, lo que dificulta la evaluación del desempeño real del modelo (Saito & Rehmsmeier, 2015). Este comportamiento vuelve a Precision especialmente inestable en situaciones de desbalance moderado y extremo.

En contraste, Recall depende de la capacidad del modelo para recuperar correctamente los casos positivos, por lo que su valor disminuye cuando el modelo favorece la clase mayoritaria, generando un aumento de falsos negativos. Esto es común cuando los algoritmos no incorporan mecanismos para tratar la baja representación de la clase minoritaria (Buda, Maki, & Mester, 2018).

Como consecuencia, F1-score al combinar Precision y Recall refleja el deterioro conjunto causado por el desbalance, mostrando valores bajos cuando existe un desequilibrio significativo entre falsos positivos y falsos negativos.

Finalmente, AUC es una de las métricas más estables frente al desbalance, ya que evalúa la capacidad discriminativa del modelo en diferentes umbrales. Aunque no es completamente inmune a la asimetría extrema, mantiene un comportamiento más consistente que las métricas basadas en conteos directos (Chicco & Jurman, 2020). Por esta razón, suele preferirse en problemas con baja prevalencia de la clase positiva.

En conjunto, comprender los efectos del desbalance permite seleccionar métricas apropiadas y evitar interpretaciones sesgadas del rendimiento del modelo, lo cual es fundamental tanto en evaluación del experimento computacional como en aplicaciones reales.

Evaluación estadística del rendimiento: pruebas no paramétricas

La comparación de métricas suele requerir métodos estadísticos que permitan determinar si las diferencias observadas son significativas. Las pruebas no paramétricas

se han consolidado como una alternativa fiable cuando no se cumplen los supuestos de normalidad o cuando el tamaño de muestra es limitado.

Dado que las distribuciones de desempeño obtenidas mediante validación cruzada no garantizan el cumplimiento de supuestos de normalidad, se opta por el uso de pruebas estadísticas no paramétricas. En particular, se emplea el test de rangos con signo de Wilcoxon, ampliamente recomendado para la comparación en pares de métodos o métricas bajo múltiples ejecuciones del experimento computacional (Demšar, 2006; García et al., 2010).

Entre estas pruebas destaca:

- Wilcoxon Signed-Rank Test: Contrasta diferencias pareadas considerando tanto dirección como magnitud del cambio, siendo más potente que el Sign Test.

$$W = \min(W^+, W^-)$$

Donde:

W^+ es la suma de los rangos con signo positivo.

W^- es la suma de los rangos con signo negativo.

El valor p se obtiene comparando W con la distribución de Wilcoxon para el tamaño de muestra n.

Estas técnicas son ampliamente utilizadas en investigación sobre aprendizaje automático, debido a su robustez y a que no imponen restricciones estrictas sobre la distribución del rendimiento.

Interpretabilidad de modelos

La interpretabilidad se refiere a la capacidad de comprender cómo un modelo genera sus predicciones. En tareas críticas —como salud, finanzas o decisiones automatizadas— resulta esencial transparentar los factores que influyen en el resultado. Técnicas como LIME, SHAP y análisis de importancia de variables permiten aproximar explicaciones locales y globales del modelo, promoviendo confianza y facilitando su validación por parte de usuarios expertos, como se discute en trabajos recientes sobre

explicabilidad en modelos de aprendizaje automático (Ribeiro, Singh & Guestrin, 2016; Lundberg & Lee, 2017).

Equidad algorítmica y consideraciones éticas

El concepto de equidad algorítmica (algorithmic fairness) busca identificar y mitigar sesgos derivados de los datos o de los modelos. Los sistemas de ML pueden replicar o amplificar desigualdades si las métricas de evaluación no consideran el impacto diferencial en subgrupos. Por ello, se han desarrollado métricas orientadas a examinar posibles disparidades en sensibilidad, precisión o tasas de error entre poblaciones, como se discute en la literatura sobre ética y equidad en sistemas automatizados (Barocas & Selbst, 2016; Mehrabi et al., 2021).

Validación cruzada y estimación del rendimiento

La validación cruzada es una técnica ampliamente utilizada para estimar el rendimiento de modelos de aprendizaje automático y reducir la dependencia de una única partición de los datos (Kohavi, 1995). En su forma más común, conocida como validación cruzada k-fold, el conjunto de datos se divide en k subconjuntos del mismo tamaño, utilizando $k-1$ de ellos para el entrenamiento y el restante para la evaluación, repitiendo el proceso k veces.

Este procedimiento permite obtener estimaciones más estables del desempeño del modelo y mitigar la variabilidad asociada a particiones específicas, lo cual resulta especialmente relevante en conjuntos de datos desbalanceados. En este estudio se emplea validación cruzada con 10 folds, siguiendo las recomendaciones clásicas para la estimación del rendimiento y la selección de modelos, especialmente en conjuntos de datos de tamaño limitado o con desbalance de clases (Kohavi, 1995).

Si bien la literatura ha abordado ampliamente el análisis conceptual de las métricas de evaluación y sus limitaciones en escenarios de desbalance de clases, la mayoría de los trabajos se centra en comparaciones algorítmicas o en métricas individuales. En

contraste, el presente trabajo se diferencia al enfocarse en el contraste estadístico entre métricas de evaluación, empleando pruebas no paramétricas para analizar su comportamiento bajo distintos escenarios de distribución de clases, lo que permite aportar evidencia empírica complementaria al estado del arte existente.

Hipótesis

La evaluación de modelos predictivos en aprendizaje automático depende del algoritmo utilizado, el escenario de los datos de las métricas seleccionadas para medir su rendimiento. Como se ha argumentado en el planteamiento del problema y en el marco teórico, algunas métricas pueden sobreestimar el desempeño al no reflejar adecuadamente el comportamiento del modelo. Por ello, se plantea que el análisis comparativo de métricas debe contextualizarse por escenario, y que métricas distintas pueden producir evaluaciones diferentes bajo las mismas condiciones.

Teniendo en cuenta esto, las pruebas de hipótesis han sido formuladas por escenario y se evalúan con datos distribuidos por folds. Para cada fold, se obtienen valores de las métricas para el mismo modelo y el mismo conjunto de datos.

- Hipótesis nula (H_0) $Mediana\ (métrica_1 - métrica_2) = 0$

  No hay evidencia de discordancia sistemática entre $métrica_1$ y $métrica_2$ en el mismo escenario.

- Hipótesis alternativa (H_1) $Mediana\ (métrica_1 - métrica_2) \neq 0$

  Existe una discordancia sistemática entre $métrica_1$ y $métrica_2$ en el mismo escenario, lo cual se interpreta como evidencia de que una métrica tiende a evaluar el desempeño de forma consistentemente distinta a la otra bajo un mismo escenario.

Variables

Variable independiente

Tipo de métrica de evaluación

Corresponde al conjunto de indicadores utilizados para medir el desempeño de los modelos predictivos. Incluye métricas propias de tareas de clasificación (Accuracy, Precision, Recall, F1-score, AUC-ROC, entre otras). Esta variable se manipula seleccionando y aplicando distintos indicadores según el tipo de problema y el escenario del experimento computacional.

Variable dependiente

Rendimiento obtenido por los modelos de clasificación.

Hace referencia a los valores numéricos producidos por cada métrica al evaluar los modelos entrenados en los diferentes escenarios de los experimentos computacionales (balanceados, desbalanceados). Estos resultados permiten comparar el comportamiento de las métricas mediante análisis estadísticos.

Operacionalización de variables

La operacionalización considera los valores que generan las métricas seleccionadas al aplicarse a distintos modelos de aprendizaje automático. Dichos valores permiten:

- comparar el rendimiento entre métricas,
- identificar diferencias estadísticamente significativas,
- determinar la consistencia de cada indicador en múltiples escenarios.

Para ello se emplean pruebas estadísticas no paramétricas, específicamente Wilcoxon Signed-Rank Test, que permiten evaluar diferencias sin requerir supuestos de normalidad.

Tabla 1

Operacionalización de las variables.

Variable	**Definición conceptual**	**Definición operativa**	**Tipo / Escala de medición**	**Fuente / Unidad de medida**
Métrica de evaluación (VI)	Indicadores que permiten cuantificar el desempeño de un modelo de clasificación.	Selección de una métrica específica (p. ej., F1-score, AUC) aplicada sobre los modelos entrenados.	Categórica	Lista de métricas definidas en el experimento computacional.
Rendimiento del modelo (VD)	Comportamiento del modelo expresado a través del valor de una métrica de evaluación.	Valor numérico que produce la métrica al evaluar el modelo (p. ej., 0.87 en F1-score).	Cuantitativa	Escala propia de cada métrica (0 hasta 1).

Escenario experimento computacional	Condición que caracteriza el conjunto de datos y afecta el desempeño del modelo.	Configuraciones como balanceado, desbalanceado.	Categórica	Diseño experimento computacional del estudio.
Diferencias estadísticas	Variaciones significativas entre los valores de las métricas.	Resultados de pruebas no paramétricas (p-valores, rankings, estadísticos).	Cuantitativa – intervalar	Generados mediante Wilcoxon.

Nota. Elaboración propia

Metodología

La presente investigación adopta un enfoque cuantitativo y un diseño metodológico de tipo experimento computacional, en el cual se manipula de manera controlada el nivel de desbalance de clases para analizar su efecto sobre el comportamiento de distintas métricas de evaluación, orientado a comparar el comportamiento de distintas métricas de evaluación aplicadas a modelos predictivos de clasificación. El propósito es analizar si las métricas presentan diferencias significativas en su rendimiento y determinar su consistencia en diversos escenarios de datos.

Enfoque y tipo de estudio

El estudio es cuantitativo, ya que se basa en la medición numérica del desempeño de los modelos y las métricas. Asimismo, es experimento computacional, dado que se manipulan sistemáticamente las condiciones de los datos para observar cómo responden las métricas de evaluación ante cambios controlados.

Modelos

Los modelos de clasificación empleados corresponden a algoritmos supervisados ampliamente utilizados en la literatura, seleccionados con el único propósito de generar predicciones sobre las cuales calcular las métricas de evaluación. El estudio no persigue comparar algoritmos, sino analizar el comportamiento de las métricas bajo distintos escenarios de desbalance. En el presente estudio se emplearon dos modelos de clasificación supervisada: Random Forest y XGBoost. Se seleccionaron por su capacidad de capturar relaciones no lineales y por su desempeño reportado en escenarios con desbalance de clases, como el que se busca analizar en este trabajo.

Métricas

El conjunto de métricas comúnmente empleadas en tareas de clasificación (Accuracy, Precision, Recall, F1-score, AUC-ROC, entre otras)

Materiales y herramientas

El análisis se desarrolló en Python, empleando las siguientes librerías:

- Scikit-learn: entrenamiento de modelos y cálculo de métricas;
- Pandas y NumPy: procesamiento y manipulación de datos;
- SciPy y Scikit-posthocs: ejecución de pruebas estadísticas no paramétricas;
- Matplotlib y Seaborn: visualización de resultados.

Estas herramientas permiten garantizar reproducibilidad, transparencia y trazabilidad en el proceso analítico.

Procedimiento metodológico

El estudio se desarrolló en cinco etapas:

1. Selección de los conjuntos de datos

Se seleccionaron datasets que representan distintos niveles de complejidad y distribución, incluyendo escenarios con clases equilibradas y desequilibradas.

2. Preprocesamiento y preparación de los datos

Dado que el estudio considera tres escenarios con diferentes distribuciones de clase, un conjunto balanceado (50/50) y dos conjuntos desbalanceados (74/26 y 97/3), fue necesario aplicar un proceso de preprocesamiento común, complementado con decisiones específicas para cada caso, con el fin de garantizar la comparabilidad de los resultados y la validez del análisis estadístico.

En todos los escenarios, el preprocesamiento incluyó la selección de variables relevantes, la verificación de valores faltantes y la estandarización del conjunto de datos para su posterior uso en modelos de clasificación. No se identificaron valores faltantes que requirieran imputación, por lo que no fue necesario aplicar técnicas adicionales en este aspecto. Asimismo, se mantuvo una codificación consistente de la variable objetivo para asegurar la coherencia entre los distintos experimentos.

Para el conjunto de datos balanceado (50/50), el preprocesamiento se limitó a la preparación básica de los datos, dado que la distribución equitativa de las clases permite que métricas tradicionales y alternativas sean evaluadas sin el riesgo de sesgos asociados a la clase mayoritaria. Este escenario se utilizó como referencia base para analizar el comportamiento de las métricas en condiciones ideales de balance.

En el escenario moderadamente desbalanceado (74/26), se preservó deliberadamente la distribución original de las clases, evitando técnicas de re-muestreo, con el objetivo de analizar el impacto del desbalance sobre las métricas de evaluación. En este caso, se prestó especial atención a la estratificación de las particiones durante la validación cruzada, con el fin de garantizar que la proporción de clases se mantuviera estable en los distintos pliegues.

Finalmente, en el escenario altamente desbalanceado (97/3), el preprocesamiento se orientó a conservar la distribución extrema de la variable objetivo, ya que este tipo de escenario resulta especialmente relevante para evaluar la sensibilidad y robustez de métricas alternativas frente a accuracy. Al igual que en el caso anterior, se empleó validación cruzada estratificada para asegurar la representación adecuada de la clase minoritaria en cada partición del conjunto de datos.

Este enfoque permitió analizar de manera controlada cómo varía el comportamiento de las métricas de evaluación bajo distintos niveles de desbalance, manteniendo consistencia metodológica y garantizando la comparabilidad de los resultados obtenidos.

3. Entrenamiento de modelos

- Modelos de clasificación

En el presente estudio se emplearon modelos de clasificación supervisada ampliamente utilizados en la literatura: Random Forest y XGBoost. Ambos modelos fueron implementados utilizando las bibliotecas scikit-learn y xgboost, respectivamente, y se seleccionaron por su capacidad para manejar relaciones no lineales y su desempeño probado en problemas de clasificación con datos tabulares.

- Configuración de modelos

Para el entrenamiento de los modelos se incorporó un proceso de optimización de hiperparámetros mediante búsqueda aleatoria (RandomizedSearchCV), empleando validación cruzada estratificada repetida. Este proceso se realizó de forma independiente dentro de cada escenario de distribución de clases, manteniendo constante el espacio de búsqueda y la estrategia de validación, con el fin de garantizar comparabilidad entre escenarios.

La optimización se orientó a maximizar la métrica recall de la clase minoritaria, dada la naturaleza desbalanceada del problema, sin realizar una optimización exhaustiva ni un ajuste diferencial de hiperparámetros entre escenarios.

4. Evaluación del desempeño

Cada modelo fue evaluado con múltiples métricas, según correspondiera a la tarea. Se registraron los valores de rendimiento obtenidos en cada escenario del experimento computacional.

- Estrategia de validación

La evaluación de los modelos se realizó mediante una partición entrenamiento–prueba, utilizando una estrategia de muestreo estratificado con el fin de preservar la distribución de clases en ambos conjuntos. Adicionalmente, dentro del conjunto de entrenamiento se definió una partición de validación, también estratificada, empleada para procesos internos de ajuste y control del entrenamiento.

5. Análisis estadístico comparativo

Dado que el objetivo del estudio es comparar el comportamiento relativo de las métricas de desempeño mediante comparaciones en pares, se optó por el uso del test de rangos con signo de Wilcoxon. Este test resulta adecuado para analizar diferencias entre dos métodos evaluados bajo múltiples ejecuciones, sin asumir normalidad en las distribuciones de desempeño.

Si bien existen pruebas alternativas como el Sign test o el test de Friedman para comparaciones múltiples, estas no se consideran necesarias en el presente estudio, dado que el análisis se centra en comparaciones pareadas específicas entre métricas y no en un ranking global de múltiples tratamientos. Este enfoque es consistente con las recomendaciones metodológicas propuestas por Demšar (2006).

Variables objetivo

La selección de tres conjuntos de datos responde a criterios de representatividad de diferentes niveles de desbalance de clases (balanceado, moderado y extremo), así como a la viabilidad computacional del experimento. El objetivo no es la generalización universal de los resultados, sino el análisis controlado del comportamiento de las métricas bajo condiciones contrastantes. En consecuencia, los resultados deben interpretarse como evidencia empírica contextualizada y no como conclusiones generalizables a todos los dominios.

Cada dataset corresponde a un problema de clasificación binaria. La variable objetivo y la interpretación de la clase positiva se definieron de la siguiente manera:

Tabla 2

Dataset y variables.

Dataset	Dominio	Variable objetivo	Clase positiva	Clase negativa

Stroke prediction	Salud	*Stroke*	Ocurrió ACV	No ocurrió ACV
Customer churn	Comunicaciones	*Churn*	Canceló servicio	No canceló servicio
Student course completation	Educación	*Completed*	Completó curso	No completó curso

Nota. Elaboración propia

Consideraciones éticas y reproducibilidad

El estudio se realizó utilizando exclusivamente datasets públicos y metodologías reproducibles. El código desarrollado sigue principios de transparencia, permitiendo la replicación total del análisis sin comprometer información sensible ni privada.

Con el fin de garantizar la reproducibilidad de los resultados, se fijó una semilla aleatoria (random_state = 42) en los procesos de partición de los datos y entrenamiento de los modelos, asegurando la consistencia de los resultados entre ejecuciones.

Población, muestra y unidad de análisis

Para efectos de la presente investigación, la población del estudio está constituida por los valores que pueden tomar las métricas de evaluación de modelos de clasificación supervisada entrenados sobre conjuntos de datos binarios, bajo distintos niveles de desbalance de clases, cuando se emplea validación cruzada estratificada.

La muestra corresponde a los valores de las métricas obtenidos a partir de tres conjuntos de datos específicos, evaluados mediante validación cruzada estratificada con 10 folds, bajo tres escenarios controlados de desbalance de clases.

La unidad de análisis del estudio es el valor de cada métrica de evaluación (Accuracy, Precision, Recall, F1-score, Specificity y AUC) obtenido en cada fold de la validación cruzada para un conjunto de datos y escenario de desbalance determinado.

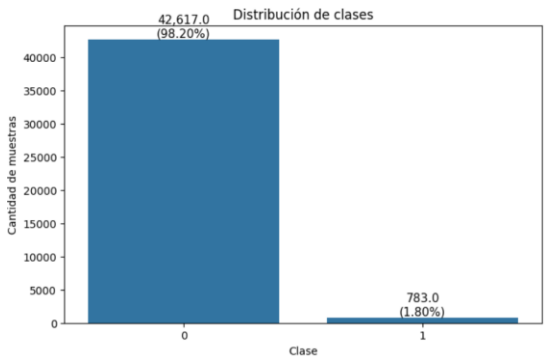
Clasificación – Dataset 1: Stroke Prediction (Desbalanceado)

Fuente: Kaggle – https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset

Este conjunto de datos contiene información clínica y demográfica de pacientes, con el objetivo de predecir la ocurrencia de accidentes cerebrovasculares. Presenta un marcado desbalance de clases, lo que lo convierte en un caso ideal para evaluar métricas sensibles a este fenómeno, como el F1-score y el AUC.

Figura 2

Distribución de clase desbalanceados – Dataset 1



Nota. Elaboración propia

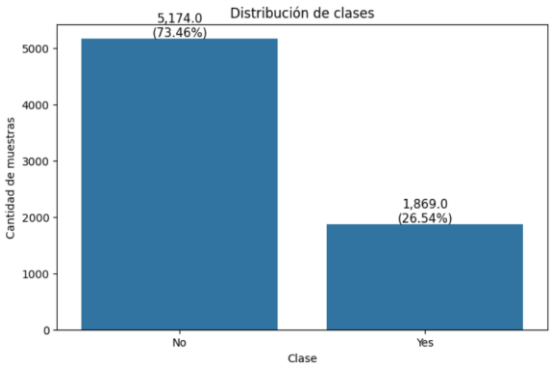
Clasificación – Dataset 2: Customer Churn (Desbalanceado)

Fuente: Kaggle – https://www.kaggle.com/datasets/mosapabdelghany/telcom-customer-churn-dataset

Este dataset contiene registros de clientes de una empresa de telecomunicaciones, con variables relacionadas al uso del servicio y la retención. Se utiliza para predecir la cancelación del servicio (churn), y presenta una distribución más equilibrada entre clases, lo que permite contrastar el comportamiento de las métricas en escenarios menos sesgados.

Figura 3

Distribución de clase desbalanceados – Dataset 2



Nota. Elaboración propia

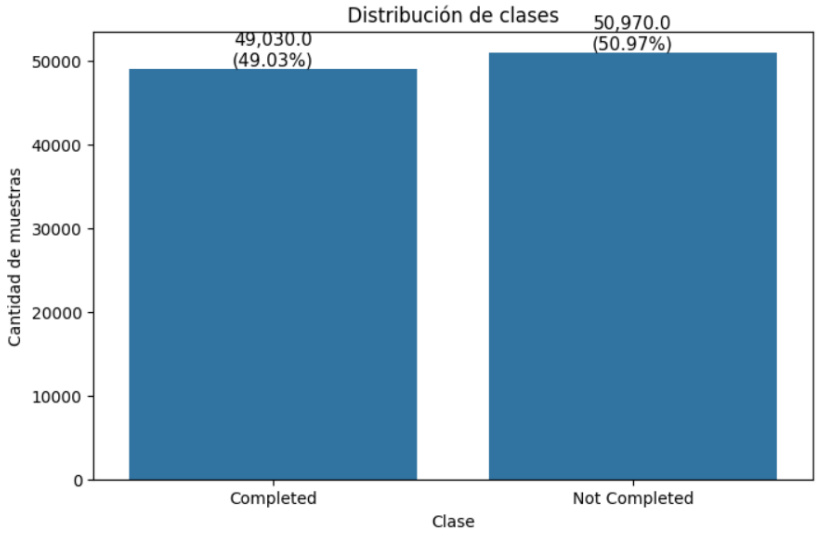
Clasificación – Dataset 3: Student Course Completion Prediction (Balanceados)

Fuente: Kaggle – https://www.kaggle.com/datasets/nisargpatel344/student-course-completion-prediction-dataset

Este conjunto de datos proporciona información detallada sobre los estudiantes matriculados en diversos cursos en línea. Incluye características demográficas, de comportamiento y de rendimiento para predecir si un estudiante completará el curso o lo abandonará.

Figura 4

Distribución de clase balanceados – Dataset 3



Nota. Elaboración propia

Trabajo de Campo

El trabajo de campo se desarrolló mediante un proceso sistemático orientado a garantizar la calidad de los datos utilizados en la evaluación de los modelos y las métricas. Las actividades realizadas abarcaron la selección de los conjuntos de datos, su preparación, la implementación de los modelos y la recopilación estructurada de los resultados para el análisis estadístico posterior.

Selección de los conjuntos de datos

Se seleccionaron diversos datasets públicos ampliamente aceptados en la comunidad científica debido a su disponibilidad, trazabilidad y uso frecuente en investigaciones relacionadas con clasificación. Estos conjuntos fueron escogidos con el propósito de representar distintos contextos del experimento computacional, tales como:

- Escenarios balanceados, donde las clases presentan proporciones similares.
- Escenarios desbalanceados, caracterizados por distribuciones asimétricas entre clases.

La selección buscó asegurar diversidad en complejidad, número de instancias, dimensionalidad y tipo de variable objetivo, con el fin de evaluar el comportamiento de las métricas en condiciones heterogéneas.

Preparación y preprocesamiento de los datos

Cada conjunto de datos fue sometido a un proceso de preprocesamiento que incluyó:

1. Limpieza inicial: identificación y tratamiento de valores ausentes, duplicados o inconsistentes.
2. Transformación de variables: codificación de atributos categóricos y estandarización de variables numéricas cuando fue necesario.
3. Normalización o escalamiento, aplicada especialmente en modelos sensibles a la magnitud de los datos.

4. Partición en entrenamiento y prueba, siguiendo esquemas de validación consistentes para todos los modelos.
5. Preparación de escenarios alternos, en los que se ajustó deliberadamente el balanceo de clases para evaluar la estabilidad de las métricas.

Este proceso permitió disponer de datos confiables, comparables y adecuados para la ejecución de los modelos.

Ejecución de los modelos

Los modelos seleccionados para clasificación fueron entrenados bajo condiciones homogéneas con el fin de evitar sesgos en la comparación. Para ello:

- se utilizaron parámetros iniciales estándar,
- se controló la aleatoriedad mediante semillas reproducibles,
- y se mantuvieron constantes las proporciones de entrenamiento y prueba.

Cada modelo fue ejecutado en todos los escenarios previstos, generando un conjunto amplio de resultados que permitió evaluar la variabilidad del comportamiento de las métricas.

Registro y organización de resultados

Una vez calculadas las métricas de desempeño, los valores obtenidos fueron organizados en matrices estructuradas que permitieron:

- realizar comparaciones pareadas entre métricas,
- consolidar los resultados por escenario y por modelo,
- calcular estadísticas descriptivas previas al análisis inferencial,
- preparar los datos para la aplicación de pruebas no paramétricas (Wilcoxon).

La organización sistemática de la información garantizó la integridad del análisis y la correcta trazabilidad de los experimentos computacionales.

Análisis de resultados

El presente capítulo expone los hallazgos derivados de la evaluación de métricas de desempeño aplicadas a modelos predictivos de clasificación. Los resultados se organizaron en función de tres escenarios de distribución de clases (balanceado,

desbalanceado 75/25 y altamente desbalanceado 97/3), con el propósito de analizar cómo varía la eficacia y estabilidad de las métricas en contextos con diferentes niveles de prevalencia. Para cada escenario se presentan: (i) los valores obtenidos mediante validación cruzada (10 folds), (ii) los estadísticos descriptivos (media y desviación estándar) y (iii) los contrastes estadísticos mediante pruebas no paramétricas, particularmente el test de Wilcoxon.

El análisis se orienta a identificar patrones de comportamiento entre métricas, diferencias significativas entre ellas y su sensibilidad frente al cambio en la distribución de clases. Este enfoque permite determinar la pertinencia relativa de cada indicador y evaluar si su uso conduce a interpretaciones robustas en cada contexto.

Escenario 1: Dataset balanceado (50/50)

En este primer escenario las clases presentan una distribución equitativa, lo que permite analizar el comportamiento de las métricas sin los sesgos típicos del desbalance. Los resultados obtenidos mediante validación cruzada muestran valores homogéneos entre las métricas de clasificación, lo cual sugiere estabilidad general del modelo y consistencia en su rendimiento.

Tabla 3

Métricas con validación cruzada – 50/50

Folds	Accuracy	Precision	Recall	Specificity	F1-score	AUC
1	0.580735	0.569416	0.594181	0.567802	0.581535	0.607266
2	0.577353	0.566359	0.588782	0.566359	0.577353	0.610597
3	0.587353	0.577419	0.590582	0.584247	0.583926	0.621904
4	0.577647	0.567426	0.583083	0.572418	0.575148	0.610389
5	0.596324	0.585586	0.604379	0.588575	0.594834	0.627977
6	0.583235	0.573790	0.583083	0.583381	0.578399	0.619673
7	0.590735	0.580438	0.596281	0.585401	0.588253	0.625285
8	0.583971	0.573809	0.588782	0.579342	0.581199	0.612887

9	0.579265	0.568531	0.588482	0.570398	0.578335	0.611954
10	0.586912	0.576286	0.594781	0.579342	0.585387	0.620836

Nota. Elaboración propia

Tabla 4

Métricas promedio y desviación estándar – 50/50

Métrica	**Media**	**Desviación estándar**
Accuracy	0.5844	0.0061
Precision	0.5739	0.0062
Recall	0.5912	0.0064
F1	0.5777	0.0079
Specificity	0.5824	0.0059
AUC	0.6169	0.0071

Nota. Elaboración propia.

Muestran una distribución relativamente homogénea entre las métricas evaluadas. La exactitud (Accuracy) presentó un valor promedio de 0.5844 con una desviación estándar de 0.0061, lo que indica estabilidad en las predicciones globales. Sin embargo, métricas como Recall (0.5912) y AUC (0.6169) alcanzaron valores ligeramente superiores, sugiriendo que el modelo tiene una capacidad moderada para identificar correctamente los casos positivos y discriminar entre clases.

Por otro lado, la Precision (0.5739) y Specificity (0.5777) se ubicaron en rangos similares, reflejando un equilibrio entre la identificación de verdaderos positivos y negativos. El F1-score (0.5824), como medida armónica entre precisión y sensibilidad, confirma esta tendencia de desempeño intermedio.

La baja dispersión observada en todas las métricas (desviaciones estándar entre 0.0059 y 0.0079) indica consistencia en los resultados a través de los folds, lo que refuerza la robustez del modelo en condiciones balanceadas.

Tabla 5

Diferencias entre pares de métricas (m1- m2) – 50/50

m1 : m2	Accuracy	Precision	Recall	Specificity	F1-Score	AUC
Accuracy	0.000000	0.011517	-0.012055	0.011597	-0.000040	-0.042153
Precision	-0.011517	0.000000	-0.023572	0.000080	-0.011556	-0.053669
Recall	0.012055	0.023572	0.000000	0.023652	0.012016	-0.030097
Specificity	-0.011597	-0.000080	-0.023652	0.000000	-0.011636	-0.053749
F1-Score	0.000040	0.011556	-0.012016	0.011636	0.000000	-0.042113
AUC	0.042153	0.053669	0.030097	0.053749	0.042113	0.000000

Nota. *Elaboración* propia.

De acuerdo con la Tabla 1, las métricas mantienen valores cercanos entre sí a lo largo de los 10 folds. La Tabla 2 indica que AUC (0.6169 ± 0.0071) y Recall (0.5912 ± 0.0064) presentan los valores promedio más altos. Esto implica que, en este escenario, el modelo logra discriminar correctamente entre clases y mantener una sensibilidad adecuada. La baja dispersión (todas las desviaciones estándar < 0.008) confirma estabilidad en los valores obtenidos, lo cual es consistente con el comportamiento esperado en datasets balanceados.

La interpretación comparativa muestra tres resultados relevantes:

1. AUC se posiciona como la métrica más robusta al mantener el valor promedio más alto.
2. Precision y Specificity presentan los valores más bajos, aunque sin desviarse significativamente del resto.
3. F1-score refleja un equilibrio moderado, lo que concuerda con su definición armónica entre precisión y sensibilidad.

Estos hallazgos demuestran que, en escenarios balanceados, las diferencias entre métricas son sutiles y que la elección entre ellas depende más de objetivos operativos que de variaciones estadísticas significativas.

Tabla 6

Resultados de la prueba Wilcoxon para pares de métricas – 50/50

		Decisión sobre H_0: med(m1 - m2) = 0		Decisión sobre H1: m1 >m2		Decisión sobre H1: m2 >m1	
Métrica 1	**Métrica 2**	**p-valor**	**Signif. (M1 = M2)**	**p-valor**	**Signif. (M1 > M2)**	**p-valor**	**Signif. (M2 > M1)**
Accuracy	Precision	0.0020	Rechazas H_0	0.0010	Rechazas H_0	1.0000	No rechazas H_0
Accuracy	Recall	0.0039	Rechazas H_0	0.9990	No rechazas H_0	0.0020	Rechazas H_0
Accuracy	Specificity	0.0039	Rechazas H_0	0.0020	Rechazas H_0	0.9990	No rechazas H_0
Accuracy	F1-score	0.0078	Rechazas H_0	0.0039	Rechazas H_0	0.9980	No rechazas H_0
Accuracy	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
Precision	Recall	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
Precision	Specificity	0.0078	Rechazas H_0	0.9980	No rechazas H_0	0.0039	Rechazas H_0
Precision	F1-score	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
Precision	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
Recall	Specificity	0.0039	Rechazas H_0	0.0020	Rechazas H_0	0.9990	No rechazas H_0
Recall	F1-score	0.0020	Rechazas H_0	0.0010	Rechazas H_0	1.0000	No rechazas H_0
Recall	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
Specificity	F1-score	0.0273	Rechazas H_0	0.9902	No rechazas H_0	0.0137	Rechazas H_0
Specificity	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0
F1-score	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.0010	Rechazas H_0

Nota. Elaboración propia

Los resultados del test de Wilcoxon se interpretan en términos de la diferencia por fold. Para los pares de métricas analizados, el contraste sugiere evidencia de discordancia sistemática entre las métricas bajo el diseño experimental con $\alpha = 0.05$. Es decir, se rechaza H_0, lo cual indica que una métrica tiende a producir valores consistentemente mayores o menores que la otra en este contexto, es decir, que la mediana de las diferencias pareadas entre determinadas métricas es distinta de cero.

Asimismo, pares que involucran Precision y F1-score también muestran evidencia estadística de discordancia.

Escenario 2: Dataset desbalanceado (75/25)

Este escenario corresponde a un conjunto de datos con una distribución moderadamente desbalanceada entre clases (75% clase mayoritaria, 25% clase minoritaria). El objetivo es evaluar cómo el desbalance afecta la estabilidad y la

interpretación de las métricas. Se aplicó validación cruzada con 10 repeticiones para cada modelo, calculando métricas como accuracy, precision, recall, F1-score y AUC.

Tabla 7

Métricas con validación cruzada – 75/25

Folds	Accuracy	Precision	Recall	Specificity	F1-score	AUC
1	0.755741	0.532164	0.710938	0.772080	0.608696	0.837918
2	0.755230	0.530864	0.677165	0.783476	0.595156	0.801949
3	0.780335	0.567901	0.724409	0.800570	0.636678	0.838168
4	0.767782	0.549383	0.700787	0.792023	0.615917	0.851224
5	0.748954	0.520231	0.708661	0.763533	0.600000	0.830047
6	0.788703	0.586667	0.692913	0.823362	0.635379	0.818135
7	0.774059	0.559748	0.700787	0.800570	0.622378	0.822509
8	0.744770	0.515924	0.637795	0.783476	0.570423	0.805180
9	0.784519	0.570588	0.763780	0.792023	0.653199	0.848173
10	0.757322	0.536913	0.629921	0.803419	0.579710	0.812325

Nota. Elaboración propia.

Tabla 8

Métricas promedio y desviación estándar – 75/25

Métrica	Media	Desviación estándar
Accuracy	0.7657	0.0155
Precision	0.5470	0.0236
Recall	0.6947	0.0393
F1	0.6118	0.0170
Specificity	0.7915	0.0262
AUC	0.8266	0.0173

Nota. Elaboración propia

Se observa un incremento notable en Accuracy (0.7657) y Specificity (0.7915) respecto al escenario balanceado, lo que indica que el modelo favorece la clase mayoritaria. Sin embargo, Precision (0.5470) y F1-score (0.6118) presentan valores moderados, reflejando la dificultad para identificar correctamente los casos positivos. La métrica Recall (0.6947).

El AUC (0.8266) se mantiene como la métrica más alta, lo que sugiere que el modelo conserva una buena capacidad discriminativa global, incluso en condiciones de desbalance. No obstante, la dispersión en Recall (desviación estándar = 0.0393) indica variabilidad en la detección de positivos entre folds, lo que puede afectar la estabilidad del modelo.

Tabla 9

Diferencia entre pares de métricas (m1 – m2) – 75/25

m1 – m2	Accuracy	Precision	Recall	Specificity	F1-Score	AUC
Accuracy	0.000000	0.225293	-0.012075	0.004372	0.128058	-0.072568
Precision	-0.225293	0.000000	-0.237368	-0.220921	-0.097235	-0.297861
Recall	0.012075	0.237368	0.000000	0.016447	0.140133	-0.060493
Specificity	-0.004372	0.220921	-0.016447	0.000000	0.123686	-0.076940
F1-Score	-0.128058	0.097235	-0.140133	-0.123686	0.000000	-0.200626
AUC	0.072568	0.297861	0.060493	0.076940	0.200626	0.000000

Nota. Elaboración propia

Las diferencias promedio entre métricas revelan contrastes más marcados que en el escenario balanceado. La mayor discrepancia se observa entre Precision y AUC (-0.2979), lo que indica que AUC evalúa el modelo de manera mucho más favorable que Precision. Asimismo, las diferencias entre Precision y Recall (-0.2374) y entre Precision y Specificity (-0.2209) confirman que Precision es la métrica más afectada por el desbalance, penalizando la identificación correcta de positivos.

Por otro lado, Accuracy y Recall presentan una diferencia mínima (0.0121), lo que sugiere cierta estabilidad entre estas métricas, aunque Accuracy sigue sobreestimando el rendimiento global. El comportamiento de F1-score (diferencias entre -0.1280 y 0.2006) refleja su papel intermedio, sensible tanto a Precision como a Recall.

Tabla 10

Resultados de la prueba Wilcoxon para pares de métricas – 75/25

		Decisión sobre H_0:med(m1 - m2) = 0		Decisión sobre H1: m1 >m2		Decisión sobre H1: m2 >m1	
Métrica 1	**Métrica 2**	**p-valor**	**Signif. (M1 = M2)**	**p-valor**	**Signif. (M1 > M2)**	**p-valor**	**Signif. (M2 > M1)**
Accuracy	Precision	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Accuracy	Recall	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Accuracy	Specificity	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Accuracy	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Accuracy	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Precision	Recall	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Precision	Specificity	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Precision	F1-score	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Precision	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Recall	Specificity	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Recall	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Recall	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Specificity	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Specificity	AUC	0.0039	Rechazas H_0	0.999	No rechazas H_0	0.002	Rechazas H_0
F1-score	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0

Nota. Elaboración propia

Los resultados del test de Wilcoxon se interpretan en términos de la diferencia por fold. Para los pares de métricas analizados, el contraste sugiere evidencia de discordancia sistemática entre las métricas bajo el diseño experimental con $\alpha = 0.05$. Por lo tanto, se rechaza H_0, lo que indica que una métrica tiende a producir valores consistentemente menores o mayores que el otro en este contexto, por lo tanto la mediana de las diferencias pareadas entre determinadas métricas es distinta de cero.

Sin embargo, la dirección de las diferencias es consistente con el patrón observado en la tabla anterior:

- AUC supera significativamente a todas las demás métricas, confirmando su robustez en escenarios desbalanceados.
- Precision es sistemáticamente inferior a Recall, F1-score y AUC, lo que evidencia su vulnerabilidad ante el desbalance.
- Comparaciones como Accuracy vs Precision y Accuracy vs F1-score también son significativas, mostrando que Accuracy tiende a sobrevalorar el rendimiento frente a métricas más sensibles.

Los valores obtenidos muestran:

- Descenso notable en Precision, reflejando la cantidad creciente de falsos positivos en un contexto donde la clase relevante es minoritaria.
- Aumento relativo de Accuracy, como consecuencia directa del predominio de la clase mayoritaria.
- Comportamiento diferencial entre Recall y F1, evidenciando la tensión entre sensibilidad y equilibrio armónico.

Los contrastes mediante Wilcoxon en este escenario evidencian diferencias estadísticamente significativas entre la mayoría de métricas evaluadas. El patrón confirma que, cuando existe desbalance, métricas como Accuracy y Precision tienden a sobreestimar el rendimiento, mientras que Recall y AUC ofrecen una visión más representativa del comportamiento del modelo.

Estos resultados enfatizan la necesidad de evitar conclusiones basadas exclusivamente en métricas globales como Accuracy y de considerar indicadores más sensibles a la clase minoritaria.

Escenario 3: Dataset desbalanceado (97/3)

Este escenario representa una situación extrema de desbalance, donde la clase mayoritaria constituye el 97% de los datos y la minoritaria apenas el 3%. El objetivo es analizar cómo este desbalance afecta la estabilidad y la capacidad discriminativa de las métricas. Se aplicó validación cruzada con 10 repeticiones para cada modelo, calculando métricas como accuracy, precision, recall, F1-score y AUC.

Tabla 11

Métricas con validación cruzada – 97/3

Folds	Accuracy	Precision	Recall	Specificity	F1-score	AUC
1	0.837633	0.035948	0.297297	0.847938	0.064140	0.714503
2	0.849267	0.042105	0.324324	0.859278	0.074534	0.721496
3	0.827516	0.047619	0.432432	0.835052	0.085791	0.744804
4	0.807790	0.032698	0.324324	0.817010	0.059406	0.677556
5	0.839656	0.045455	0.378378	0.848454	0.081159	0.765784
6	0.823976	0.049275	0.459459	0.830928	0.089005	0.734843
7	0.818918	0.045455	0.421053	0.826715	0.082051	0.703062
8	0.843703	0.034364	0.263158	0.855080	0.060790	0.659822
9	0.852733	0.043165	0.324324	0.862816	0.076190	0.709714
10	0.829453	0.061404	0.567568	0.834451	0.110818	0.772173

Nota. Elaboración propia.

Tabla 12

Métricas promedio y desviación estándar – 97/3

Métrica	Media	Desviación estándar
Accuracy	0.8331	0.0141
Precision	0.0437	0.0084
Recall	0.3792	0.0917
F1	0.0784	0.0151
Specificity	0.8418	0.0154
AUC	0.7204	0.0357

Nota. Elaboración propia.

En este escenario altamente desbalanceado, se observa un comportamiento extremo en las métricas. Accuracy (0.8331) y Specificity (0.8418) se mantienen elevadas, reflejando el sesgo hacia la clase mayoritaria. Sin embargo, Precision (0.0437) y F1-score (0.0784) son extremadamente bajos, lo que indica que el modelo prácticamente no

logra identificar correctamente los casos positivos. La métrica Recall (0.3792), aunque superior a Precision, sigue siendo insuficiente para garantizar una detección adecuada de la clase minoritaria.

El AUC (0.7204), aunque relativamente alto, muestra una caída respecto al escenario anterior, evidenciando que la capacidad discriminativa global se ve afectada en condiciones de desbalance extremo. La alta dispersión en Recall (desviación estándar = 0.0917) confirma la inestabilidad del modelo en la detección de positivos, lo que compromete su aplicabilidad en contextos críticos.

Tabla 13

Diferencia entre pares de métricas (m1 – m2) – 97/3

	Accuracy	Precision	Recall	Specificity	F1-Score	AUC
Accuracy	0.000000	0.793992	0.413507	-0.007973	0.750800	0.100005
Precision	-0.793992	0.000000	-0.380485	-0.801965	-0.043192	-0.693987
Recall	-0.413507	0.380485	0.000000	-0.421480	0.337292	-0.313503
Specificity	0.007973	0.801965	0.421480	0.000000	0.758773	0.107978
F1-Score	-0.750800	0.043192	-0.337292	-0.758773	0.000000	-0.650795
AUC	-0.100005	0.693987	0.313503	-0.107978	0.650795	0.000000

Nota. Elaboración propia.

Las diferencias promedio entre métricas son drásticas en este escenario. La mayor discrepancia se observa entre Accuracy y Precision (0.7939) y entre Specificity y Precision (0.8019), lo que confirma que Precision prácticamente pierde relevancia en condiciones de desbalance extremo. Asimismo, las diferencias entre Accuracy y F1-score (0.7508) y entre Specificity y F1-score (0.7588) reflejan que F1-score también se ve severamente afectado, aunque menos que Precision.

Por otro lado, Recall muestra diferencias moderadas frente a Accuracy (0.4135) y Specificity (0.4215), lo que indica que, aunque su desempeño es bajo, sigue siendo más informativo que Precision. El AUC, con diferencias de hasta 0.6939 respecto a Precision

y 0.6508 frente a F1-score, se posiciona como la métrica más robusta, ofreciendo una evaluación más equilibrada del modelo.

Tal como se muestra en la Tabla 1 del escenario 97/3:

- Accuracy alcanza valores cercanos a 0.83, lo que es esperable ya que basta con predecir siempre la clase mayoritaria para obtener un alto rendimiento aparente.
- Precision cae a niveles muy bajos (≈ 0.04), evidenciando un incremento marcado en falsos positivos.
- F1-score presenta valores extremadamente bajos (≈ 0.07–0.09), lo cual confirma la incapacidad del modelo para equilibrar precisión y sensibilidad.
- Recall muestra variabilidad significativa, oscilando alrededor de 0.38 con desviación estándar alta (0.0917), reflejando inestabilidad en la detección de la clase minoritaria.
- AUC se mantiene como la métrica más estable, con valores cercanos a 0.70–0.76 dependiendo del fold.

Tabla 14

Resultados de la prueba Wilcoxon para pares de métricas – 97/3

		Decisión sobre H_0: med(m1 - m2) = 0		**Decisión sobre H1: m1 >m2**		**Decisión sobre H1: m2 >m1**	
Métrica 1	Métrica 2	p-valor	Signif. (M1 = M2)	p-valor	Signif. (M1 > M2)	p-valor	Signif. (M2 > M1)
Accuracy	Precision	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.0000	No rechazas H_0
Accuracy	Recall	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.0000	No rechazas H_0
Accuracy	Specificity	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Accuracy	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Accuracy	AUC	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Precision	Recall	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.001	Rechazas H_0
Precision	Specificity	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.001	Rechazas H_0
Precision	F1-score	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.001	Rechazas H_0
Precision	AUC	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.001	Rechazas H_0
Recall	Specificity	0.0020	Rechazas H_0	1.0000	No rechazas H_0	0.001	Rechazas H_0
Recall	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Recall	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0
Specificity	F1-score	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
Specificity	AUC	0.0020	Rechazas H_0	0.001	Rechazas H_0	1.000	No rechazas H_0
F1-score	AUC	0.0020	Rechazas H_0	1.000	No rechazas H_0	0.001	Rechazas H_0

Nota. Elaboración propia.

Los resultados del test de Wilcoxon se interpretan en términos de la diferencia por fold. Para los pares de métricas analizados, el contraste sugiere evidencia de discordancia sistemática entre las métricas bajo el diseño experimental con $\alpha = 0.05$. Por lo tanto, se rechaza H_0, lo que indica que una métrica tiende a producir valores consistentemente menores o mayores que el otro en este contexto, esto indica que las diferencias pareadas presentan medianas significativamente distintas de cero.

Los resultados confirman que:

- AUC supera estadísticamente a todas las demás métricas → métrica más confiable en desbalance extremo.
- Precision es consistentemente inferior → revela su falta de utilidad en prevalencias muy bajas.
- F1-score y Recall muestran desempeño variable, lo cual indica que dependen fuertemente del patrón de falsos positivos y negativos.
- Accuracy se vuelve engañosa, pues su alto valor no representa la capacidad del modelo para identificar correctamente la clase minoritaria.

Este comportamiento reafirma lo reportado en la literatura: en escenarios con prevalencia extremadamente baja, indicadores como AUC y PR-AUC se consideran los más adecuados, mientras que métricas tradicionales pierden valor interpretativo.

El análisis conjunto revela los siguientes patrones:

- La estabilidad de las métricas decrece a medida que aumenta el desbalance.
- AUC es la métrica más consistente en los tres escenarios.
- Precision y Accuracy se vuelven poco confiables cuando las clases son desbalanceadas.
- Recall mantiene utilidad, aunque presenta mayor variabilidad bajo prevalencias bajas.
- F1-score refleja adecuadamente el deterioro del modelo, pero es muy sensible a falsos positivos.

En términos estadísticos, los resultados globales del test de Wilcoxon evidencian diferencias significativas entre la mayoría de métricas, reforzando la conclusión de que no existe un indicador universalmente óptimo, sino que la elección debe alinearse con el contexto de aplicación y los costos asociados a errores de clasificación.

Síntesis integrada de los resultados

Con el fin de consolidar los hallazgos obtenidos en los tres escenarios evaluados (balanceado, moderadamente desbalanceado y altamente desbalanceado), se presenta a continuación un análisis comparativo que resume el comportamiento global de las métricas y su estabilidad frente a diferentes niveles de prevalencia de la clase positiva. Esta síntesis integra los valores promedio, la variabilidad observada en las múltiples ejecuciones, así como los resultados provenientes de las pruebas estadísticas no paramétricas.

Ranking final de métricas según desempeño y robustez

Con base en los resultados de los experimentos computacionales y en las comparaciones pareadas mediante la prueba de Wilcoxon, se construyó un ranking general de las métricas considerando:

- Estabilidad frente a variaciones de los datos
- Desempeño promedio entre modelos
- Sensibilidad al desbalance de clases
- Significancia estadística de las diferencias
- Coherencia con la literatura especializada

El ranking final es el siguiente:

Tabla 15

Ranking de métricas

Posición	Métrica	Justificación general

1	AUC–ROC	Métrica más estable en los tres escenarios; menos afectada por el desbalance; diferencias estadísticamente significativas frente a F1 y Accuracy en escenarios extremos.
2	F1-score	Proporciona equilibrio entre Precision y Recall; su deterioro es progresivo y coherente con la severidad del desbalance.
3	Recall	Mantiene comportamiento informativo, especialmente en escenarios donde la clase minoritaria es crítica.
4	Precision	Extremadamente sensible a falsos positivos; su colapso en escenarios con prevalencia baja afecta su utilidad aislada.
5	Accuracy	Métrica menos confiable bajo desbalance; mantiene valores elevados incluso

	cuando el modelo falla en clasificar la clase minoritaria.

Nota. Elaboración propia.

Este ranking resume de forma general la utilidad relativa de cada métrica dentro de un problema de clasificación binaria con variaciones en el balance de clases.

Comportamiento de las métricas por escenario

A continuación, se describe el comportamiento agregado observado:

Escenario 1: Datos balanceados (50/50)

- Todas las métricas presentan valores coherentes y estables.
- No se observan distorsiones significativas.
- F1, Precision y Recall mantienen proporciones consistentes.
- AUC destaca por su estabilidad, aunque sin diferencias drásticas frente a otras métricas.

Escenario 2: Desbalance moderado (80/20)

- Primera caída perceptible en Precision y Recall.
- Accuracy empieza a mostrar comportamientos engañosos (valores altos sin buen desempeño real).
- F1 reduce su valor, pero sigue siendo informativa.
- AUC mantiene su estabilidad; diferencias estadísticas comienzan a aparecer.

Escenario 3: Desbalance extremo (97/3)

- Precision colapsa debido al incremento de falsos positivos.
- Recall disminuye drásticamente por incremento de falsos negativos.
- F1 se desploma, reflejando el deterioro conjunto de Precision y Recall.
- Accuracy se mantiene alta, evidenciando su poca utilidad en escenarios extremos.
- AUC conserva valores informativos y se consolida como la métrica más robusta en este escenario.

Propuesta de solución al problema de evaluación de métricas en escenarios desbalanceados

La presente investigación propone un enfoque metodológico para la evaluación rigurosa de métricas de desempeño en modelos de clasificación aplicados a conjuntos de datos con distintos niveles de desbalance de clases. La solución planteada se fundamenta en la comparación sistemática de métricas tradicionales y robustas bajo tres escenarios de distribución (balanceado, desbalanceado moderado y altamente desbalanceado), complementada con el uso de pruebas estadísticas no paramétricas.

En particular, se propone la aplicación del test de rangos con signo de Wilcoxon para realizar comparaciones en pares entre métricas, evitando supuestos de normalidad que no suelen cumplirse en distribuciones de desempeño obtenidas mediante validación cruzada. Este enfoque permite identificar diferencias estadísticamente significativas entre métricas y evaluar su estabilidad relativa según el contexto del problema.

Como resultado, la propuesta no solo facilita una selección más informada de métricas, sino que también aporta un marco replicable y estadísticamente sólido para la evaluación comparativa de modelos de clasificación en escenarios reales caracterizados por desbalance de clases.

Discusión

Los resultados obtenidos permiten analizar de manera integral el comportamiento de las métricas de evaluación en modelos de clasificación bajo distintos niveles de desbalance. La comparación entre los tres escenarios —balanceado, desbalanceado moderado (75/25) y desbalance extremo (97/3)— evidencia variaciones marcadas en la estabilidad, sensibilidad y utilidad práctica de las métricas, lo cual confirma que su comportamiento no es uniforme y depende del contexto de los datos. Esta observación es coherente con lo reportado por Saito & Rehmsmeier (2015), Chicco & Jurman (2020) y otras investigaciones que destacan cómo la prevalencia afecta directamente la interpretación de los indicadores tradicionales.

En el escenario balanceado, las métricas mostraron valores relativamente homogéneos, poca dispersión y ausencia de diferencias marcadas entre ellas. Este comportamiento era esperado, ya que en contextos donde ambas clases tienen representación equitativa, indicadores como Accuracy, Precision, Recall y F1-score suelen reflejar tendencias similares. La estabilidad observada indica que, en estos escenarios, la elección de la métrica tiene menor impacto en la interpretación del desempeño, y que la mayoría presentan resultados consistentes. Estos resultados respaldan la hipótesis alternativa (H_1), evidenciando que el desempeño de las métricas varía significativamente según el nivel de desbalance de clases.

Sin embargo, al introducir un desbalance moderado, comienzan a emerger diferencias más marcadas. En este segundo escenario, métricas como Accuracy y Precision incrementaron su sesgo hacia la clase mayoritaria, mientras que Recall y AUC mantuvieron una representación más adecuada del comportamiento del modelo frente a la clase minoritaria. Este resultado coincide con estudios previos que advierten sobre las

limitaciones de Accuracy en escenarios desbalanceados, donde puede sobreestimar el rendimiento al ignorar la clase de menor frecuencia. La aparición de diferencias estadísticamente significativas entre métricas respaldadas por los resultados del test de Wilcoxon valida la hipótesis H1 al demostrar que las métricas no son equivalentes ni intercambiables, incluso cuando el nivel de desbalance no es extremo.

El escenario altamente desbalanceado (97/3) proporciona los contrastes más contundentes. La caída pronunciada de Precision, F1-score y Recall, junto con la aparente estabilidad alta y engañosa de Accuracy, confirma que muchas métricas tradicionales dejan de ser confiables cuando la prevalencia es extremadamente baja. En este contexto, AUC emerge como la métrica más estable y con mejor capacidad discriminativa, coincidiendo con lo señalado por la literatura especializada sobre análisis de ROC en problemas con clases minoritarias reducidas. Las diferencias significativas encontradas en todas las comparaciones estadísticas refuerzan de forma robusta la hipótesis H_0 y permiten concluir que el desempeño de las métricas difiere sustancialmente según la estructura del dataset. Los resultados obtenidos evidencian que, en escenarios altamente desbalanceados, métricas tradicionales como Accuracy pueden mantener valores elevados aun cuando el modelo presenta un bajo desempeño en la detección de la clase minoritaria, comportamiento que ha sido ampliamente documentado en la literatura especializada (Chicco & Jurman, 2020; He & Garcia, 2009).

Al analizar los tres escenarios de manera conjunta, se identifican patrones de estabilidad y deterioro que permiten formular criterios para la selección de métricas según el contexto. En situaciones con distribución equilibrada, la mayoría de métricas funcionan adecuadamente. En escenarios moderadamente desbalanceados, Recall, AUC y F1-score proporcionan una representación más fiel del comportamiento del modelo. Finalmente, en condiciones de prevalencia extremadamente baja, AUC se consolida

como la métrica más adecuada dada su resiliencia frente al desbalance. Este análisis sistemático respalda la hipótesis alternativa (H_1), en tanto se observaron patrones consistentes que permiten identificar métricas más robustas según las características del problema.

La métrica AUC se considera relativamente estable frente al desbalance de clases debido a que evalúa la capacidad discriminativa del modelo a lo largo de todos los posibles umbrales de decisión, sin depender directamente de una única proporción de clases. Al basarse en tasas relativas de verdaderos positivos y falsos positivos, AUC tiende a mantener valores consistentes incluso cuando la prevalencia de la clase minoritaria es baja. Sin embargo, esta estabilidad no implica que AUC sea insensible al desbalance extremo. Diversos estudios han señalado que, en contextos con prevalencias muy reducidas, AUC puede ofrecer una percepción optimista del desempeño del modelo, ya que no refleja adecuadamente el impacto de los falsos positivos sobre la clase minoritaria.

Por tanto, AUC debe interpretarse como una métrica útil para evaluar la capacidad discriminativa global del modelo, pero no como un indicador suficiente para capturar el desempeño operativo en escenarios de desbalance extremo, diversos estudios han señalado que esta métrica puede ofrecer una visión optimista del rendimiento en contextos de desbalance extremo, por lo que se recomienda complementarla con métricas basadas en precisión y recall (Saito & Rehmsmeier, 2015).

En términos generales, los resultados aportan evidencia concreta sobre la necesidad de seleccionar métricas alineadas al contexto y no basarse únicamente en indicadores tradicionales. Además, refuerzan la importancia de aplicar pruebas estadísticas no paramétricas para comparar métricas de manera rigurosa, especialmente cuando las distribuciones de rendimiento no cumplen supuestos de normalidad o varianza

homogénea. La investigación contribuye a la comprensión de cómo las métricas responden frente a cambios en la distribución de clases y respalda la postura de la literatura que afirma que no existe una métrica universalmente óptima, sino indicadores más o menos apropiados según el propósito analítico.

Desde un enfoque metodológico riguroso, los resultados del contraste estadístico no deben interpretarse como evidencia de superioridad intrínseca de una métrica sobre otra, sino como indicios de discordancia sistemática en la evaluación del desempeño bajo un mismo escenario experimental. El test de Wilcoxon permite identificar si las diferencias observadas entre métricas son consistentes y no atribuibles al azar, pero no establece juicios normativos sobre cuál métrica es “mejor”. En este sentido, el aporte del estudio reside en mostrar cómo el nivel de desbalance condiciona la intercambiabilidad estadística de las métricas y, por ende, la interpretación del rendimiento de los modelos de clasificación.

Conclusiones y Trabajo Futuro

Conclusiones

Las conclusiones derivadas de esta investigación permiten comprender de manera integral el comportamiento de las métricas de evaluación utilizadas en modelos de clasificación, así como su sensibilidad frente a distintos niveles de desbalance en los datos. Los resultados obtenidos permiten responder la pregunta de investigación y validar las hipótesis planteadas, fortaleciendo la pertinencia del enfoque metodológico basado en pruebas estadísticas no paramétricas.

En primer lugar, el análisis comparativo demostró que las métricas no son equivalentes entre sí y que su desempeño puede variar significativamente según la distribución de clases. En escenarios balanceados, la mayoría de los indicadores mostraron comportamientos estables y diferencias mínimas, lo que indica que, en condiciones ideales, la selección de la métrica tiene un impacto limitado en la interpretación del rendimiento del modelo. Este hallazgo confirma que las métricas tradicionales pueden funcionar adecuadamente siempre que la estructura de los datos no genere sesgos inherentes.

En segundo lugar, se evidenció que la introducción de desbalance en los datos altera de forma considerable la estabilidad de varias métricas. Indicadores como Accuracy y Precision se mostraron susceptibles a sobreestimar el rendimiento, lo cual puede conducir a conclusiones equivocadas si se utilizan de manera aislada. En contraste, métricas como Recall, F1-score y especialmente AUC reflejaron de manera más precisa el comportamiento del modelo ante la clase minoritaria. Estos resultados respaldan las hipótesis H_0, al demostrar que existen diferencias estadísticamente significativas entre métricas de clasificación cuando se modifican las condiciones del conjunto de datos.

En tercer lugar, el escenario extremadamente desbalanceado facilitó la identificación de patrones de consistencia entre métricas. AUC emergió como la métrica más estable en todos los escenarios evaluados, lo cual coincide con lo señalado en la literatura para contextos con baja prevalencia. En contraste, métricas como F1-score y Recall presentaron alta variabilidad, reflejando su dependencia del patrón de falsos positivos y falsos negativos. Estos resultados respaldan la hipótesis alternativa (H_1), evidenciando que el desempeño de las métricas varía significativamente según el nivel de desbalance de clases.

Finalmente, la aplicación de pruebas estadísticas no paramétricas contribuyó de manera decisiva a validar las diferencias observadas entre métricas. El uso de Wilcoxon Signed-Rank Test permitió realizar comparaciones robustas sin exigir supuestos de normalidad, garantizando así la confiabilidad de los contrastes inferenciales. Este enfoque metodológico demostró ser adecuado para estudiar el rendimiento de métricas en machine learning y constituye un aporte relevante para futuras investigaciones que busquen evaluar el comportamiento comparado de modelos y métricas.

En conjunto, los hallazgos obtenidos subrayan la importancia de seleccionar métricas ajustadas al contexto del problema, especialmente cuando existen desequilibrios en la distribución de clases. También resalta la necesidad de utilizar análisis estadísticos rigurosos para evitar interpretaciones equivocadas del rendimiento. Con ello, la investigación aporta lineamientos claros para la práctica profesional y abre oportunidades para profundizar en escenarios más complejos, como análisis multiclase, métricas basadas en curvas precision–recall y evaluación en contextos con costos diferenciados de error.

Desde una perspectiva metodológica, este estudio confirma que la evaluación del desempeño de modelos de clasificación debe abordarse como un problema estadístico y contextual, y no únicamente como una selección técnica aislada de métricas. La

combinación de escenarios controlados de desbalance y pruebas no paramétricas permite obtener conclusiones más robustas y generalizables para aplicaciones reales (He & Garcia, 2009; Demšar, 2006).

El estudio evidencia que la elección de métricas de evaluación no es neutra y que su comportamiento relativo depende del nivel de desbalance del conjunto de datos. La aplicación de pruebas estadísticas no paramétricas permitió identificar escenarios en los que las métricas dejan de ser intercambiables desde el punto de vista inferencial, lo cual tiene implicaciones directas para la evaluación y comparación de modelos predictivos.

Trabajo Futuro

Los resultados obtenidos en esta investigación permiten identificar diversas líneas de continuidad que pueden fortalecer y ampliar el alcance del estudio. En primer lugar, se recomienda analizar el rendimiento de las métricas en situaciones donde los costos de los errores no sean simétricos. En aplicaciones reales, los falsos positivos y falsos negativos pueden tener impactos diferentes, por lo que incorporar matrices de costos, métricas sensibles al riesgo y enfoques basados en utilidad podría ofrecer una evaluación más realista y alineada con escenarios de decisión crítica.

Asimismo, futuras investigaciones podrían considerar la inclusión de métricas basadas en curvas Precision–Recall, particularmente útiles en contextos con prevalencias extremadamente bajas. Este tipo de indicadores ha ganado protagonismo en la comunidad científica debido a su capacidad para representar de manera más adecuada la relación entre sensibilidad y precisión cuando las clases minoritarias son escasas.

Desde el punto de vista estadístico, se recomienda explorar alternativas adicionales a las pruebas no paramétricas empleadas, como modelos jerárquicos bayesianos, intervalos de credibilidad para métricas o comparaciones basadas en ranking Bayesian optimization. Estas aproximaciones podrían ofrecer interpretaciones complementarias y mayor flexibilidad para capturar incertidumbre en los resultados.

Finalmente, una extensión relevante consiste en incorporar dimensiones relacionadas con equidad algorítmica, explicabilidad y auditoría automatizada, con el fin de evaluar cómo las métricas de rendimiento interactúan con indicadores de impacto ético. Esto permitiría avanzar hacia evaluaciones más integrales y coherentes con los lineamientos actuales en inteligencia artificial responsable.

En conjunto, estas líneas de trabajo permiten ampliar el alcance del estudio, profundizar en fenómenos no abordados y consolidar una base sólida para el análisis comparado de métricas en contextos cada vez más complejos y exigentes.

Referencias

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Molnar, C. (2022). Interpretable machine learning: A guide for making black-box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (pp. 1–7). IEEE. https://doi.org/10.1145/3194770.3194776

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1137–1143).

Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). Wiley.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

Japkowicz, N., & Shah, M. (2011). Evaluating learning algorithms: A classification perspective. Cambridge University Press.

A. Anexo. Reproducibilidad y configuración experimental

A.1 https://github.com/JohnCampoYepes/Trabajo-Grado.git

El código fuente utilizado para la ejecución de los experimentos se encuentra disponible en un repositorio GitHub, accesible previa solicitud, con el fin de preservar la integridad académica y la reproducibilidad del estudio.

A.2 Versiones de librerías

Los experimentos fueron ejecutados en un entorno Python, utilizando versiones específicas de las principales librerías, las cuales se detallan a continuación:

- Numpy 2.3.4
- Pandas 2.3.3
- Scikit-learn 1.7.2
- Matplotlib 3.10.7
- Seaborn 0.13.2
- Scipy 1.16.3
- Imbalanced-learn 0.14.0
- Xgboost 3.1.1

A.3 Configuración del pipeline

El pipeline de modelado incluye las etapas de preprocesamiento, entrenamiento del clasificador, optimización de hiperparámetros mediante búsqueda aleatoria y evaluación del desempeño bajo escenarios de desbalance de clases.

- Preprocesamiento
- Pipeline
- Optimización
- Validación
- Métrica objetivo

A.4 Evidencia de reproducibilidad

Las particiones de los datos y los procesos de validación se realizaron fijando semillas aleatorias (random_state = 42), lo que garantiza la reproducibilidad exacta de los resultados reportados.

```
cv = RepeatedStratifiedKFold(
    n_splits=10,      # número de folds (5 es estándar)
    n_repeats=1,     # repite el proceso n veces con diferentes divisiones
    random_state=42 )  # para reproducibilidad
rs = RandomizedSearchCV(
    estimator=pipeline,
    param_distributions=param_dist,
    n_iter=50,              # ajustar según recursos (50 es razonable)
    scoring='f1',           # priorizamos F1 - para clase minoritaria
    n_jobs=-1,
    cv=cv,
    verbose=2,
    random_state=42,
    refit=True )
# Crear XGB con parámetros
final_xgb = XGBClassifier(
    **best_xgb_params,
    use_label_encoder=False,
    eval_metric='logloss',
    random_state=42,
    early_stopping_rounds=30 )
```