

# SOLUCIONES PARA CONJUNTOS DE DATOS CON CLASES DESBALANCEADAS COMBINADAS CON MODELOS DE APRENDIZAJE AUTOMÁTICO ENSAMBLADOS

## SOLUTIONS FOR IMBALANCED DATASETS COMBINED WITH ENSEMBLE MACHINE LEARNING MODELS

MENESES OSPINA MARÍA ISABELLA<sup>a</sup>, RAMÍREZ SIERRA YULY ANDREA<sup>b</sup>, LAMOS  
DÍAZ HENRY<sup>c</sup>

**RESUMEN:** La tarea de clasificación en el aprendizaje automático implica predecir una etiqueta de clase para cada instancia, basándose en los patrones descubiertos durante la fase de entrenamiento del modelo, para automatizar su asignación en nuevas observaciones. No obstante, surge el *problema de desbalance de clases*, originado por tendencias de distribución sesgada. Este fenómeno se da cuando una clase está representada por un amplio número de elementos, en comparación con los elementos de las demás clases, lo que llevaría a que probablemente los modelos de aprendizaje automático tengan un desempeño deficiente durante su fase de validación, evidenciado en baja precisión e incapacidad de generalización. Por tanto, mediante una revisión de literatura se ha identificado que una de las soluciones más efectivas son los modelos ensamblados que utilizan estrategias de Bagging, Boosting o Stacking combinadas con soluciones enfocadas en el preprocesamiento del conjunto de datos. Entre estas estrategias, se destaca el uso de la técnica de sobremuestreo SMOTE, que ha demostrado aumentar el desempeño del modelo.

**PALABRAS CLAVE:** Aprendizaje automático, Problema de desbalance de clases, Datos desequilibrados, Técnicas de remuestreo, Clasificadores ensamblados.

**ABSTRACT:** The classification task in machine learning involves predicting a class label for each instance based on patterns discovered during the model's training phase to automate its assignment in new observations. However, the *class imbalance problem* arises due to skewed distribution trends. This phenomenon occurs when one class is represented by a large number of elements compared to elements in other classes, which would likely lead to machine learning models performing poorly during their validation phase, evidenced by low accuracy and a lack of generalization ability. Therefore, through a literature review, it has been identified that one of the most effective solutions is the use of ensemble models employing Bagging, Boosting, or Stacking strategies combined with data preprocessing-focused solutions. Among these strategies, the use of the SMOTE oversampling technique stands out, as it has demonstrated an improvement in model

---

<sup>a</sup>(Est.) en Ingeniería Industrial. Escuela de Estudios Industriales y Empresariales. Universidad Industrial de Santander.

<sup>b</sup>(M.Sc) en Ingeniería Industrial. Profesora cátedra. Escuela de Estudios Industriales y Empresariales. Universidad Industrial de Santander.

<sup>c</sup>(PhD) en Física - Matemática. Profesor titular. Escuela de Estudios Industriales y Empresariales. Universidad Industrial de Santander

performance.

**KEYWORDS:** Machine learning, Class imbalanced problem, Imbalanced data, Resampling techniques, Classifier ensembles.

## 1. INTRODUCCIÓN

La clasificación es una de las tareas más importantes para el aprendizaje automático (Tarekegn et al., 2021), y su aplicabilidad se extiende a diversos campos, incluyendo diagnósticos médicos, predicción de crisis financieras (Lin et al., 2017), reconocimiento de emociones (Tarekegn et al., 2021) y detección de emergencias (Jo and Japkowicz, 2004). Sin embargo, esta tarea se ve afectada por un fenómeno que se manifiesta en dichos entornos, en el cual, por las características inherentes del conjunto de datos desequilibrado, los datos tienden a agruparse en clases mayoritarias o minoritarias. Este fenómeno se conoce en literatura como *problema de desbalance de clases*, en donde “una clase puede estar representada por un amplio número de elementos, mientras que la otra parte se encuentra representada por unos pocos elementos” (Lin et al., 2017). Además, este problema a menudo implica un solapamiento significativo entre clases (Denil and Trappenberg, 2010), lo que puede dar lugar a la fragmentación de la clase minoritaria (Jo and Japkowicz, 2004).

Según Lin et al. (2017), existen diversos enfoques para dar solución al problema de desbalance de clases, los cuales se engloban en soluciones a nivel de datos, algoritmo, sensibilidad al costo y clasificadores ensamblados (p.18). No obstante, según Galar et al. (2012), la combinación de métodos de pre-procesamiento de datos con modelos de aprendizaje automático para la tarea de clasificación posee mejores rendimientos que otros métodos. En este sentido, a través de una revisión de literatura se busca identificar enfoques que dan solución a los conjuntos de datos con clases desbalanceadas a partir de modelos de aprendizaje automático ensamblados. Por tanto, en este documento se presenta la metodología utilizada para la revisión de literatura, después se detalla la sección de resultados y finalmente se concluye sobre los aspectos relevantes del tema desarrollado.

## 2. METODOLOGÍA

Para llevar a cabo un análisis específico de los enfoques que dan solución a los conjuntos de datos desbalanceados a través de estrategias de remuestreo y modelos de aprendizaje automático ensamblados, se realizó la búsqueda de artículos científicos publicados en la base de datos SCOPUS utilizando la siguiente ecuación de búsqueda: TITLE-ABS-KEY ('class-imbalanced' OR 'class-imbalance' OR 'imbalanc\* data' OR 'imbalanc\* problem' OR 'imbalanc\* classification' OR 'imbalanc\* class distribution' OR 'unbalanced data') AND TITLE-ABS-KEY ('resampling techniques' OR resampling OR oversampling OR undersampling) AND TITLE-ABS-KEY ('machine learning' OR 'supervised learning') AND TITLE-ABS-KEY ('classification algorithm\*' OR classifier OR 'classification technique') AND TITLE-ABS-KEY ('assembled classifier\*' OR 'classifier ensemble\*' OR 'ensembles of classifier\*' OR 'ensemble model\*' OR 'ensemble method'). A partir de esta ecuación, se encuentran 86 documentos, a los cuales se le aplican criterios de inclusión como periodo de publicación desde 2013 a 2023 y limitando el idioma a inglés. Además, se excluyen documentos tales como las revisiones de conferencia y las revisiones, para un total de 72 documentos; después, se revisa el título y resumen de los documentos en donde se selecciona 54 y finalmente se eligen 35 documentos a

Después se realiza el análisis bibliométrico y se identifican los hallazgos destacables en los documentos seleccionados.

Es importante señalar que ‘*classifier ensemble*’ es el término más relevante y se encuentra estrechamente relacionado con otros conceptos, como ‘*bagging*’ o ‘*boosting*’, que corresponden a métodos de aprendizaje automático ensamblados. Además, se encontraron términos como ‘*resampling*’, ‘*oversampling*’ y ‘*SMOTE*’, que están relacionados con técnicas de remuestreo de datos.

A través de los últimos años, el término ‘*classifier ensemble*’, ha evolucionado hacia un concepto más amplio conocido como ‘*ensemble learning*’. Al principio se parte de la idea de combinar predicciones individuales para mejorar el desempeño del modelo en comparación con un solo clasificador. Sin embargo, la idea de *ensemble learning* migra a un concepto más amplio que incluye tanto modelos de clasificación como de regresión, e incluye también técnicas adicionales, como ‘*feature extraction*’, ‘*feature selection*’ y ‘*k-means*’, empleadas en la fase de procesamiento de datos. Finalmente, las aplicaciones más destacadas de este enfoque se encuentran en contextos diversos, como la evaluación crediticia y la identificación de enfermedades como la diabetes mellitus o el cáncer de mama.

### 3.2. Revisión de literatura

Para abordar de manera efectiva la tarea de clasificación cuando un modelo predictivo se enfrenta al problema de desequilibrio de clases, es esencial no ceñirse únicamente a la aplicación de algoritmos de aprendizaje automático. Si bien la mayoría de estos algoritmos, algunos tales como los modelos de árboles de decisión (CART), las redes neuronales o el algoritmo de K-vecinos más cercanos (KNN), han demostrado un desempeño superior al lograr tasas de error promedio más bajas en comparación con técnicas estadísticas tradicionales, como el modelo Probit (Galindo and Tamayo, 2000). Estos algoritmos o clasificadores al enfrentarse a un conjunto de datos con desequilibrio de clases grave o con datos poco frecuentes, no son capaces de discriminar adecuadamente entre las clases mayoritarias y las clases minoritarias. Como resultado, el clasificador tiende a etiquetar a casi todos los casos como pertenecientes a la clase mayoritaria, lo que puede llevar a una métrica de precisión engañosamente alta (Hasanin et al., 2019), a pesar de un rendimiento real deficiente en términos de clasificación. Por lo tanto, se hace necesario explorar y analizar los diversos enfoques existentes destinados a tratar la complejidad del problema y mejorar el desempeño de los modelos de aprendizaje automático orientados a la clasificación.

Dos de los enfoques ampliamente utilizados son las técnicas de muestreo y los modelos de aprendizaje automático ensamblado (Guo et al., 2019). Sin embargo, los modelos ensamblados se destacan en comparación con el muestreo de los datos e incluso con algoritmos de clasificación o el proceso de selección de características; ya que, son capaces de combinar múltiples técnicas de muestreo de datos y/o algoritmos para la optimización de los clasificadores, y así, aliviar indirectamente el efecto del problema de desequilibrio de datos (Quan et al., 2022).

Las técnicas de muestreo, también conocidas como soluciones a nivel de datos (Lin et al., 2017), se centran en la etapa de preprocesamiento de los datos (Inan et al., 2021). Estas técnicas se consideran métodos externos, ya que involucran la creación de un conjunto de entrenamiento equilibrado mediante la reducción de la clase mayoritaria o el incremento de la clase minoritaria (Sun et al., 2018). Se dividen en dos categorías principales: submuestreo y sobremuestreo, en donde las formas más simples de aplicación son el submuestreo aleatorio (RUS) y el sobremuestreo aleatorio (ROS), respectivamente (Mwangi et al., 2022). Sin embargo, existe un enfoque adicional que se deriva de las técnicas de muestreo y combinan tanto el submuestreo como el sobremuestreo, a este se le denomina muestreo híbrido (Lin and Nguyen, 2020). En una investigación acerca de la calidad del agua, se llevó a cabo una comparación del desempeño de siete modelos de aprendizaje automá-

tico utilizando tanto la técnica de ROS como el RUS, además de explorar el enfoque híbrido ROS-RUS. Los resultados revelaron que la combinación del método híbrido, en combinación con el clasificador ensamblado Random Forest, condujo a mejoras significativas en términos de precisión, especificidad y la puntuación F-1; y adicionalmente la técnica ROS generalmente tuvo un mejor desempeño, pero con una ventaja mínima, seguido por el RUS (Malek et al., 2023). Sin embargo, es importante destacar que estas técnicas presentan limitaciones (Honnurappa and Raghavendra, 2021), como la pérdida de información en la clase mayoritaria o al sobreajuste de la clase minoritaria (Mwangi et al., 2022), lo que puede dar como resultado predicciones incorrectas debido a la alta varianza generada por la naturaleza aleatoria de estos métodos (Ng et al., 2017). Por otra parte, los modelos de aprendizaje automático ensamblados se aplican después de las técnicas de preprocesamiento, en la fase de implementación de los modelos. En esta etapa, se realizan predicciones de clases en función de la técnica de ensamblado empleada (Balasubramanian et al., 2020). Estos métodos de aprendizaje ensamblado se dividen principalmente en tres categorías: Bagging, Boosting y Stacking. Bagging combina las predicciones a través de votación, Boosting se basa en que el rendimiento de los modelos anteriores influye en los nuevos modelos, mientras que Stacking implica la combinación de modelos de diferentes tipos (Pouriyeh et al., 2017). En un estudio para la detección de fraudes en tarjetas de crédito en un conjunto altamente desequilibrado con un número total de casos de fraude de 492 de un total de 284,807 transacciones, se evaluaron dos clasificadores individuales: KNN y regresión logística (LR), junto con dos modelos ensamblados: el clasificador Bagging con árboles de decisión como clasificador base y Random Forest (RF). Además, se llevaron a cabo experimentos utilizando dos métodos tradicionales de sobremuestreo, como SMOTE y ADASYN. Los resultados revelaron que los modelos ensamblados presentaron un desempeño promedio superior en comparación con los modelos individuales. Además, se observó que es posible mejorar significativamente el rendimiento del modelo ensamblado al combinarlo estratégicamente con técnicas tradicionales de sobremuestreo. En particular, se encontró que la aplicación de RF después de equilibrar los datos con SMOTE condujo a un rendimiento razonablemente bueno y superó a varias combinaciones posibles de enfoques (Mondal et al., 2021).

Finalmente, una de las métricas de desempeño más importante tanto en aplicaciones médicas como en aplicaciones generales, especialmente cuando existe el problema de clases desbalanceadas, es el Recall, también conocido como recuperación o sensibilidad. Esta métrica permite monitorear el número de falsos positivos con el objetivo de reducirlos y, por tanto, mejorar el desempeño del modelo ensamblado (Balasubramanian et al., 2020). Sin embargo, dado que la precisión no puede capturar completamente el desempeño del clasificador, se recomienda utilizar la métrica de la media geométrica armónica, que resulta útil en situaciones de desequilibrio de clases. Otras medidas de desempeño valiosas incluyen la sensibilidad y la precisión. Además, en algunas investigaciones, se recomienda explorar nuevas formas de mejorar la especificidad y la puntuación del área bajo la curva (AUC) de la curva de característica operativa del receptor (ROC) (Mwangi et al., 2022) el cual permite observar la forma como un modelo discrimina entre las clases teniendo en cuenta umbrales de probabilidad.

## 4. CONCLUSIONES

Para abordar eficazmente el desequilibrio de clases en problemas de clasificación, es esencial reconocer que la aplicación de algoritmos de aprendizaje automático de clasificación por sí solos no son suficientes. Aunque muchos algoritmos, como los modelos de árboles de decisión o redes neuronales, pueden obtener un buen

desempeño en comparación con enfoques estadísticos tradicionales, luchan en situaciones de desequilibrio de clases, lo que lleva a tasas de error engañosamente bajas. Para abordar este problema, se han desarrollado dos enfoques principales: técnicas de muestreo y modelos de aprendizaje automático ensamblado. Los modelos ensamblados, que utilizan métodos como Bagging, Boosting o Stacking, se aplican después de implementar técnicas de sobremuestreo tradicionales como SMOTE en la etapa de preprocesamiento y han demostrado mejorar significativamente el desempeño de los modelos que analizan en conjuntos de datos desequilibrados. Por otra parte, el Random Forest se destaca como uno de los modelos de aprendizaje automático ensamblados más ampliamente utilizado para abordar el desequilibrio en diversas áreas de aplicación, generalmente demostrando un desempeño superior en comparación con los modelos individuales de aprendizaje automático. Además, métricas de desempeño específicas, como Recall y la media geométrica armónica, permiten una evaluación adecuada del desempeño del modelo, lo que facilita comparaciones entre modelos de aprendizaje automático, ya sean ensamblados o individuales. También es posible llevar a cabo comparaciones con diversas combinaciones de técnicas de remuestreo de datos para los casos de desequilibrio de clases. En este contexto, estudios recomiendan explorar estrategias para mejorar tanto la especificidad como la puntuación AUC en investigaciones relacionadas con este desafío.

## Referencias

- Balasubramanian, S., Kashyap, R., Cvn, S. T., and Anuradha, M. (2020). Hybrid prediction model for type-2 diabetes with class imbalance. In *Proceedings of the 2020 IEEE International Conference on Machine Learning and Applied Network Technologies, ICMLANT 2020*.
- Denil, M. and Trappenberg, T. (2010). Overlap versus imbalance. In *Advances in artificial intelligence, 23rd canadian conference on artificial intelligence*, pages 220–231.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(4):463–498.
- Galindo, J. and Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. In *Computational Economics*, volume 15.
- Guo, H., Diao, X., and Liu, H. (2019). Improving undersampling-based ensemble with rotation forest for imbalanced problem. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(2):1371–1386.
- Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., and Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, 6(1).
- Honnurappa, S. and Raghavendra, B. K. (2021). A highly robust heterogeneous deep ensemble assisted multi-feature learning model for diabetic mellitus prediction. *International Journal of Performability Engineering*, 17(11):926–937.
- Inan, M. S. K., Hasan, R., and Alam, F. I. (2021). A hybrid probabilistic ensemble based extreme gradient boosting approach for breast cancer diagnosis. In *2021 IEEE 11th Annual Computing*

- and Communication Workshop and Conference, *CCWC 2021*, pages 1029–1035.
- Jo, T. and Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1):40–49.
- Lin, Wei-Chao. and Tsai, C.-F., Hu, Y.-H., and Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, pages 17–26.
- Lin, H. I. and Nguyen, M. C. (2020). Boosting minority class prediction on imbalanced point cloud data. *Applied Sciences (Switzerland)*, 10(3).
- Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Md Nasir, S. A., Shaadan, N., and Indratno, S. W. (2023). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(1):598–608.
- Mondal, I. A., Haque, M. E., Hassan, A. M., and Shatabda, S. (2021). Handling imbalanced data for credit card fraud detection. In *24th International Conference on Computer and Information Technology, ICCIT 2021*.
- Mwangi, P. I., Nderu, L., Mutanu, L., and Mwirereri, D. G. (2022). Hybrid ensemble model for handling class imbalance problem in big data analytics. In *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022*.
- Ng, W. W. Y., Zhang, Y., and Zhang, J. (2017). Bsmboost for imbalanced pattern classification problems. [https://doi.org/10.0/Linux-x86\\_64](https://doi.org/10.0/Linux-x86_64).
- Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., and Gutierrez, J. (2017). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *Proceedings - IEEE Symposium on Computers and Communications*, pages 204–207.
- Quan, D., Feng, W., Dauphin, G., Wang, X., Huang, W., and Xing, M. (2022). A novel double ensemble algorithm for the classification of multi-class imbalanced hyperspectral data. *Remote Sensing*, 14(15).
- Sun, B., Chen, H., Wang, J., and Xie, H. (2018). Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. *Frontiers of Computer Science*, 12(2):331–350.
- Tarekegn, A., Giacobini, M., and Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:2.