Generación de instancias sintéticas guiada por dependencias entre variables en problemas de clasificación desbalanceada

Francisco-Blas Martínez-Pareja[1] and Esther-Lydia Silva-Ramírez[1]

Department of Computer Science and Engineering, University of Cádiz, Avda. Universidad de Cádiz 10, Puerto Real, Cádiz, 11519, Spain
blas.martinezpa@alum.uca.es, esther.silva@uca.es
Fecha de inicio: 25 de noviembre de 2025

Resumen El desequilibrio de clases dificulta el entrenamiento de modelos supervisados en dominios críticos como la ciberseguridad o el diagnóstico médico. Aunque los métodos de sobremuestreo sintético como SMOTE mitigan este problema, su dependencia exclusiva de criterios de proximidad geométrica local suele ignorar las relaciones estadísticas entre atributos, generando instancias inconsistentes. Esta investigación propone un nuevo enfoque de sobremuestreo guiado explícitamente por medidas de dependencia entre variables, asumiendo que la incorporación de información estructural del espacio de datos produce muestras más coherentes con la distribución real. Avalado por resultados preliminares prometedores en un Trabajo de Fin de Máster previo, el plan de tesis doctoral abarca desde el diseño formal de estos mecanismos basados en relevancia estadística hasta su evaluación exhaustiva en modelos de diversa complejidad y dominios reales altamente desbalanceados. La principal aportación del trabajo consiste en introducir mecanismos de generación sintética que incorporan información sobre la dependencia entre variables, ajustando el proceso de interpolación en función de dicha estructura estadística. Con ello se busca reducir la generación de instancias poco representativas y mejorar la capacidad de generalización de los modelos en escenarios con desequilibrio de clases.

Keywords: Aprendizaje desbalanceado · SMOTE · Generación sintética de datos · Dependencia entre variables.

1. Introducción

El aprendizaje automático se ha consolidado como una herramienta fundamental para la toma de decisiones en múltiples dominios donde los datos presentan un alto grado de complejidad. Sin embargo, uno de los principales retos en este ámbito es el desequilibrio de clases, una situación en la que las categorías de un conjunto de datos no están representadas equitativamente, existiendo una clase mayoritaria con una presencia predominante y una clase minoritaria con

una representación significativamente reducida. Este problema resulta especialmente crítico en aplicaciones donde los eventos de interés son poco frecuentes, como la detección de anomalías, el diagnóstico médico o el fraude financiero.

Entre las estrategias más utilizadas para abordar este problema se encuentran las técnicas de sobremuestreo, que buscan equilibrar la distribución de las clases mediante generación de nuevas instancias de la clase minoritaria. Métodos como SMOTE han demostrado ser eficaces en este contexto [1,2], aunque presentan limitaciones al basarse exclusivamente en criterios de proximidad local, ignorando las relaciones existentes entre variables.

La hipótesis de esta investigación es que la incorporación de información sobre la dependencia entre variables en el proceso de generación de datos sintéticos permite mejorar la calidad de las instancias generadas y, en consecuencia, el rendimiento de los modelos de clasificación en problemas con desequilibrio de clases. Esta hipótesis ha sido previamente explorada en un estudio inicial realizado en el marco de un Trabajo de Fin de Máster, cuyos resultados experimentales muestran mejoras en el rendimiento de los modelos, lo que respalda la viabilidad del enfoque propuesto para esta tesis doctoral.

En este contexto, el objetivo principal de la tesis es desarrollar nuevas estrategias de sobremuestreo que tengan en cuenta la relevancia y relación entre atributos, con el fin de generar datos sintéticos más representativos de la estructura real de los datos.

Para alcanzar este objetivo se plantean los siguientes subobjetivos:

- Analizar en profundidad las distintas técnicas de sobremuestreo existentes y sus limitaciones.
- Diseñar métodos que integren medidas de dependencia entre variables en el proceso de generación de datos.
- Evaluar el impacto de estas técnicas en distintos modelos de aprendizaje automático.
- Validar su aplicabilidad en diferentes dominios en los que es frecuente un gran desequilibrio de clases.

La aportación original de esta investigación radica en el desarrollo de un enfoque de sobremuestreo que integra explícitamente información sobre la dependencia entre variables para guiar la generación de instancias sintéticas, incorporando así conocimiento estructural del espacio de datos en el proceso de aprendizaje. Este enfoque permitirá reducir la generación de muestras inconsistentes o poco representativas, y mejorar la capacidad de los modelos para identificar correctamente la clase minoritaria en escenarios altamente desbalanceados.

2. Antecedentes

El problema del desequilibrio de clases ha sido ampliamente estudiado en el ámbito del aprendizaje automático [3], dando lugar a múltiples estrategias orientadas a la mitigación del sesgo hacia la clase mayoritaria. Entre estas estrategias, el remuestreo del conjunto de datos constituye una de las aproximaciones más

utilizadas, ya que permite modificar la distribución de las clases sin alterar los algoritmos de aprendizaje.

Las técnicas de remuestreo se dividen principalmente en dos categorías: *undersampling* y *oversampling*. *Undersampling* consiste en reducir el número de instancias de la clase mayoritaria con el objetivo de equilibrar el conjunto de datos. Si bien esta técnica permite disminuir el coste computacional del entrenamiento, puede implicar la pérdida de información relevante. Por otro lado, *oversampling* incrementa el número de ejemplos de la clase minoritaria, generalmente mediante la replicación de instancias existentes. Aunque esta aproximación preserva la información disponible, puede inducir sobreajuste debido a la redundancia de los datos generados.

Con el objetivo de superar estas limitaciones, se han desarrollado métodos de sobremuestreo más avanzados basados en la generación de instancias sintéticas. Entre ellos, destaca SMOTE (*Synthetic Minority Over-sampling Technique*) [4], que genera nuevas muestras interpolando entre instancias minoritarias y sus vecinos más cercanos en el espacio de características. Este enfoque mejora la representatividad de la clase minoritaria sin introducir duplicados exactos, favoreciendo la capacidad de generalización de los modelos.

A partir de SMOTE han surgido numerosas variantes que buscan adaptarse a distintos escenarios. Borderline-SMOTE [5] focaliza la generación de muestras en las regiones cercanas a la frontera de decisión, donde la clasificación resulta más compleja. Este enfoque permite mejorar la sensibilidad del modelo en zonas críticas del espacio de características. Por su parte, ADASYN [6] introduce un mecanismo adaptativo que ajusta la cantidad de instancias sintéticas en función de la dificultad de aprendizaje de cada ejemplo, priorizando aquellas regiones donde la clase minoritaria está peor representada.

Otras propuestas han tratado de mejorar la calidad de las muestras generadas mediante la incorporación de información adicional. Safe-Level-SMOTE [7] introduce el concepto de «nivel de seguridad»de cada instancia, definido en función de la proporción de vecinos minoritarios, con el fin de evitar la generación de ejemplos en regiones dominadas por la clase mayoritaria. En una línea similar, SMOTE-IPF [8] combina SMOTE con un proceso de filtrado iterativo que elimina instancias ruidosas o mal clasificadas, mejorando la calidad del conjunto de datos resultante.

Más recientemente, han surgido enfoques basados en aprendizaje profundo, como Deep SMOTE [9], que transforman los datos a un espacio latente mediante redes neuronales antes de generar las muestras sintéticas. Este tipo de métodos permite capturar estructuras más complejas en datos de alta dimensionalidad, aunque a costa de un incremento significativo en la complejidad computacional. En la línea orientada a la eficiencia computacional, Fast-SMOTE [10] propone generar instancias sintéticas utilizando la moda, el mínimo y el máximo de cada atributo calculados de forma global sobre la clase minoritaria, evitando así la búsqueda de vecinos más cercanos y reduciendo la complejidad temporal. Sin embargo, al igual que SMOTE clásico, este enfoque ignora las dependen-

cias estadísticas entre variables, limitando la representatividad estructural de las muestras generadas.

A pesar de los avances introducidos por estas técnicas, la mayoría de los métodos de sobremuestreo continúan basándose en criterios de proximidad local, sin considerar de forma explícita la dependencia estadística entre variables. Esta limitación puede dar lugar a la generación de instancias que no respetan las relaciones intrínsecas del conjunto de datos, introduciendo ruido y reduciendo la eficacia de los modelos de clasificación [3].

En este contexto, surge la necesidad de desarrollar nuevas estrategias que integren información sobre la relación entre atributos en el proceso de generación de datos sintéticos. La presente investigación se sitúa en esta línea, proponiendo un enfoque que incorpora medidas de dependencia entre variables con el objetivo de generar instancias más coherentes y representativas de la estructura real de los datos.

3. Propuesta de investigación

La presente investigación aborda el problema del desequilibrio de clases en tareas de clasificación mediante una extensión de los métodos de sobremuestreo sintético. En particular, se propone un enfoque que, además de considerar la proximidad local entre instancias, incorpore información sobre la dependencia entre variables durante el proceso de generación de nuevos ejemplos.

La hipótesis central es que los métodos tradicionales, como SMOTE y sus variantes, al basarse exclusivamente en criterios geométricos, pueden generar instancias sintéticas que no respetan las relaciones existentes entre atributos. Esto puede dar lugar a ejemplos poco representativos o inconsistentes con la estructura interna de los datos. En este trabajo se plantea que la incorporación de medidas de dependencia entre variables puede contribuir a generar muestras más coherentes y, por tanto, mejorar el rendimiento en problemas desbalanceados.

3.1. Formalización del problema

Sea un conjunto de datos de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^{n}$, donde cada instancia es un vector $x_i \in R^d$, siendo d la dimensionalidad del espacio de características, y $y_i \in \{C_{min}, C_{may}\}$ representa la etiqueta de clase. En escenarios de desequilibrio severo, la cardinalidad de la clase minoritaria es tal que $|S_{min}| \ll |S_{may}|$.

Siguiendo el esquema de SMOTE, una nueva instancia sintética se genera a partir de un ejemplo $x_i \in S_{min}$ y uno de sus vecinos cercanos $x_j \in S_{min}$:

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \tag{1}$$

donde $\lambda \in [0, 1]$. Este mecanismo asume que la interpolación lineal entre instancias cercanas preserva la validez de los datos, sin tener en cuenta el papel que desempeña cada variable en la representación de las instancias.

3.2. Incorporación de dependencias entre variables

La propuesta de investigación plantea extender el mecanismo clásico de generación de instancias sintéticas de SMOTE incorporando información sobre la dependencia entre variables. Mientras que SMOTE aplica una interpolación uniforme en todas las dimensiones del espacio de características, la aproximación propuesta introduce un comportamiento diferenciado para cada atributo en función de su relevancia estadística. De esta manera, la interpolación se realizará de forma dependiente de cada atributo. Para ello, se define un vector de pesos:

$$W = (w_1, w_2, \ldots, w_d) \tag{2}$$

donde cada peso w_f representa el grado de dependencia o relevancia del atributo f, estimado mediante distintas estrategias. De este modo, la generación de una nueva instancia se define componente a componente como:

$$x_{new} = x_i + \lambda_f \cdot (x_j - x_i) \tag{3}$$

donde:

$$\lambda_f = g(w_f) \tag{4}$$

y $g(\cdot)$ representa una función de ajuste que controla la variación permitida en cada dimensión en función de la relevancia del atributo.

Bajo esta formulación, los atributos con mayor dependencia estadística presentarán una menor variabilidad durante la generación sintética, favoreciendo la preservación de las relaciones estructurales presentes en los datos originales. Por el contrario, las variables menos relevantes permitirán una mayor flexibilidad en la interpolación, manteniendo la diversidad de las nuevas instancias generadas. El proceso completo de generación de instancias, desde el análisis inicial de dependencias hasta la obtención de un conjunto de datos equilibrado, se describe de forma esquemática en la Figura 1.

3.3. Beneficios esperados

Se espera que esta estrategia permita reducir la generación de instancias sintéticas inconsistentes, especialmente en aquellas regiones del espacio de características donde las relaciones entre variables desempeñan un papel determinante en la definición de las instancias. Al incorporar información sobre dependencias entre atributos, el proceso de sobremuestreo puede evitar la creación de ejemplos que, aunque válidos desde un punto de vista geométrico, resulten poco plausibles desde una perspectiva estadística. Esto podría traducirse en una mejor delimitación de la frontera de decisión entre clases y en una menor superposición artificial entre la clase minoritaria y la mayoritaria.

Como consecuencia, los modelos entrenados sobre estos conjuntos de datos enriquecidos podrían mejorar su capacidad de generalización en escenarios desequilibrados, no solo en términos de rendimiento global, sino especialmente en su habilidad para identificar correctamente instancias de la clase minoritaria.

En particular, se espera una reducción en el número de falsos negativos, lo que resulta crítico en muchas aplicaciones donde la detección de la clase minoritaria es prioritaria. La evaluación del enfoque propuesto se llevará a cabo mediante métricas específicamente adecuadas para problemas de desequilibrio de clases [3], tales como F1-score, G-Mean y el área bajo la curva Precision-Recall (AUPRC). Estas métricas permitirán capturar de manera más precisa el equilibrio entre precisión y exhaustividad en la clase minoritaria, así como el comportamiento del modelo en escenarios donde la distribución de clases es altamente asimétrica. Además, se considerará el análisis comparativo frente a métodos clásicos de sobremuestreo, con el objetivo de cuantificar el impacto real de la incorporación de dependencias entre variables en el proceso de generación sintética.

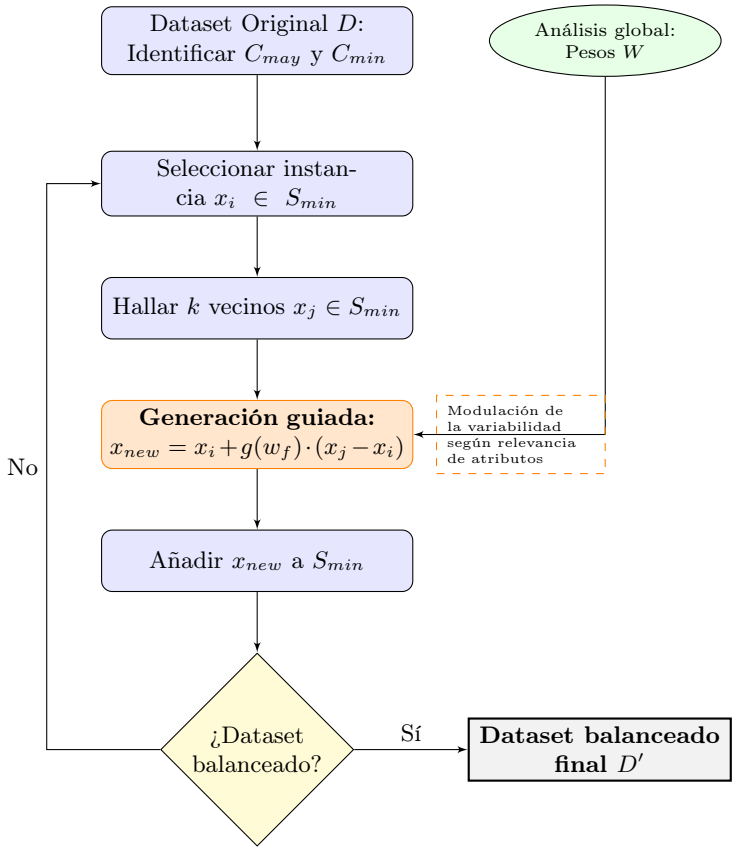


Figura 1. Diagrama de flujo del algoritmo de sobremuestreo propuesto.

4. Metodología y plan de trabajo

4.1. Enfoque metodológico

La presente investigación sigue un enfoque experimental en el ámbito de aprendizaje automático, centrado en el diseño y validación de nuevas técnicas de preprocesado de datos para problemas de clasificación desbalanceada. Con el fin de estructurar el proceso de trabajo de forma sistemática, se adoptará la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), un modelo ampliamente utilizado que organiza el ciclo de vida de un proyecto de minería de datos en fases iterativas. Este enfoque permitirá abordar el problema de manera progresiva, facilitando la validación continua de las hipótesis planteadas y la mejora del método propuesto a partir de la evidencia empírica obtenida.

De forma general, CRISP-DM se compone de las siguientes fases:

1. **Comprensión del problema**: En esta fase se analizará en profundidad el problema del desequilibrio de clases, con especial atención a sus efectos en modelos de clasificación supervisada. Se estudiarán las limitaciones de los métodos de sobremuestreo existentes, como SMOTE y sus variantes, identificando como principal limitación la ausencia de consideración explícita de las dependencias entre variables. Esta fase permitirá justificar la necesidad de un enfoque que incorpore información estructural del espacio de datos.
2. **Comprensión de los datos**: Se realizará una recopilación y un análisis exploratorio de los conjuntos de datos utilizados, evaluando aspectos como el nivel de desequilibrio, la dimensionalidad y la distribución de las variables. En esta fase también se estudiarán las relaciones entre atributos mediante medidas estadísticas con el objetivo de identificar patrones de dependencia relevantes para la futura generación de instancias sintéticas.
3. **Preparación de los datos**: En esta fase se llevará a cabo la transformación y adecuación de los datos para su posterior utilización en los modelos de aprendizaje automático. Esto incluye el preprocesamiento de las variables, aplicando técnicas de escalado para asegurar que todas las características se encuentren en rangos comparables y evitar que aquellas con mayor magnitud dominen el proceso de aprendizaje. Asimismo, se realizará la codificación de variables categóricas mediante técnicas como *one-hot encoding*, permitiendo su integración en los algoritmos de clasificación.
4. **Modelado**: Se entrenarán distintos modelos de clasificación supervisada utilizando los conjuntos de datos preprocesados con los diferentes métodos de sobremuestreo. El objetivo no es únicamente optimizar el rendimiento predictivo, sino analizar cómo la calidad de los datos sintéticos afecta al comportamiento de modelos de distinta complejidad, desde clasificadores tradicionales hasta modelos basados en ensembles, así como modelos de aprendizaje profundo.
5. **Evaluación**: Se realizará una evaluación exhaustiva de los modelos mediante métricas adecuadas para problemas desbalanceados. Además, se llevará a

cabo un análisis comparativo entre los métodos propuestos y las técnicas de referencia, con el objetivo de determinar si la incorporación de dependencias entre variables mejora de forma consistente la calidad del aprendizaje.

6. **Despliegue**: Aunque el enfoque principal de esta investigación es experimental, esta fase se interpreta desde una perspectiva de transferencia del conocimiento. Los resultados obtenidos se analizarán en términos de aplicabilidad en dominios reales, como salud o ciberseguridad, donde la generación de datos sintéticos fiables puede tener un impacto directo en sistemas de soporte a la decisión.

En el contexto de esta investigación, estas fases se abordarán de forma iterativa, priorizando la preparación de los datos para introducir el componente innovador del método, y la posterior fase de evaluación para su correspondiente validación.

4.2. Plan de trabajo

El desarrollo de la tesis se estructurará en distintas fases de investigación, cada una de ellas compuesta por un conjunto de tareas específicas orientadas al cumplimiento de los objetivos planteados. Esta organización permitirá abordar el problema de forma progresiva y facilitar el seguimiento de los hitos alcanzados durante el desarrollo de la investigación.

- **Fase 1: Análisis del estado del arte y definición del problema**
  Tareas:
  - Revisión bibliográfica sobre aprendizaje desbalanceado.
  - Estudio de técnicas de sobremuestreo basadas en SMOTE y sus variantes.
  - Análisis de enfoques recientes relacionados con generación de datos sintéticos.
  - Identificación de limitaciones asociadas a la ausencia de dependencias entre variables.
  - Definición de la hipótesis y objetivos de investigación.

  Hito: Identificación del vacío de investigación y formulación de la hipótesis de trabajo.

- **Fase 2: Diseño e implementación de los métodos propuestos**
  Tareas:
  - Definición formal de los métodos de sobremuestreo propuestos.
  - Selección de medidas de dependencia entre variables.
  - Diseño de los mecanismos de generación de instancias sintéticas.
  - Implementación de los algoritmos propuestos.
  - Validación inicial del funcionamiento de los métodos.

  Hito: Especificación formal e implementación funcional de los algoritmos propuestos.

- **Fase 3: Experimentación**
  Tareas:
  - Selección de conjuntos de datos con distintos niveles de desequilibrio.
  - Preparación y preprocesamiento de los datos.
  - Configuración de modelos de aprendizaje automático.
  - Comparación experimental con técnicas de referencia.
  - Evaluación mediante métricas específicas para clases desbalanceadas.

  Hito: Obtención de resultados experimentales en múltiples escenarios de evaluación.

- **Fase 4: Análisis de resultados**
  Tareas:
  - Análisis estadístico de los resultados obtenidos.
  - Comparación del comportamiento de los distintos métodos.
  - Evaluación del impacto de las dependencias entre variables.
  - Identificación de fortalezas y limitaciones del enfoque propuesto.

  Hito: Obtención de conclusiones sobre la efectividad de los métodos propuestos.

- **Fase 5: Redacción y difusión de resultados**
  Tareas:
  - Redacción de artículos científicos.
  - Elaboración de informes técnicos y documentación experimental.
  - Redacción de la memoria de tesis doctoral.
  - Preparación de presentaciones y difusión de resultados.

  Hito: Finalización de la tesis doctoral y publicación de resultados científicos.

Para garantizar la viabilidad del proyecto, las fases y tareas detalladas anteriormente se han organizado en un cronograma trianual (2026-2028). Esta planificación, desglosada por trimestres (Q1-Q4), contempla el solapamiento necesario entre las etapas de implementación y experimentación, así como la naturaleza transversal de la labor de difusión científica. En la Figura 2 se presenta gráficamente dicha distribución temporal.

El desarrollo de esta tesis doctoral se realiza bajo la dirección de la Dra. Esther-Lydia Silva-Ramírez, siendo Francisco-Blas Martínez-Pareja el doctorando responsable de su ejecución, con dedicación a tiempo completo. Conforme al cronograma trianual (2026-2028) descrito en la Figura 2, se estima la defensa de la tesis para finales de 2028 o principios de 2029.

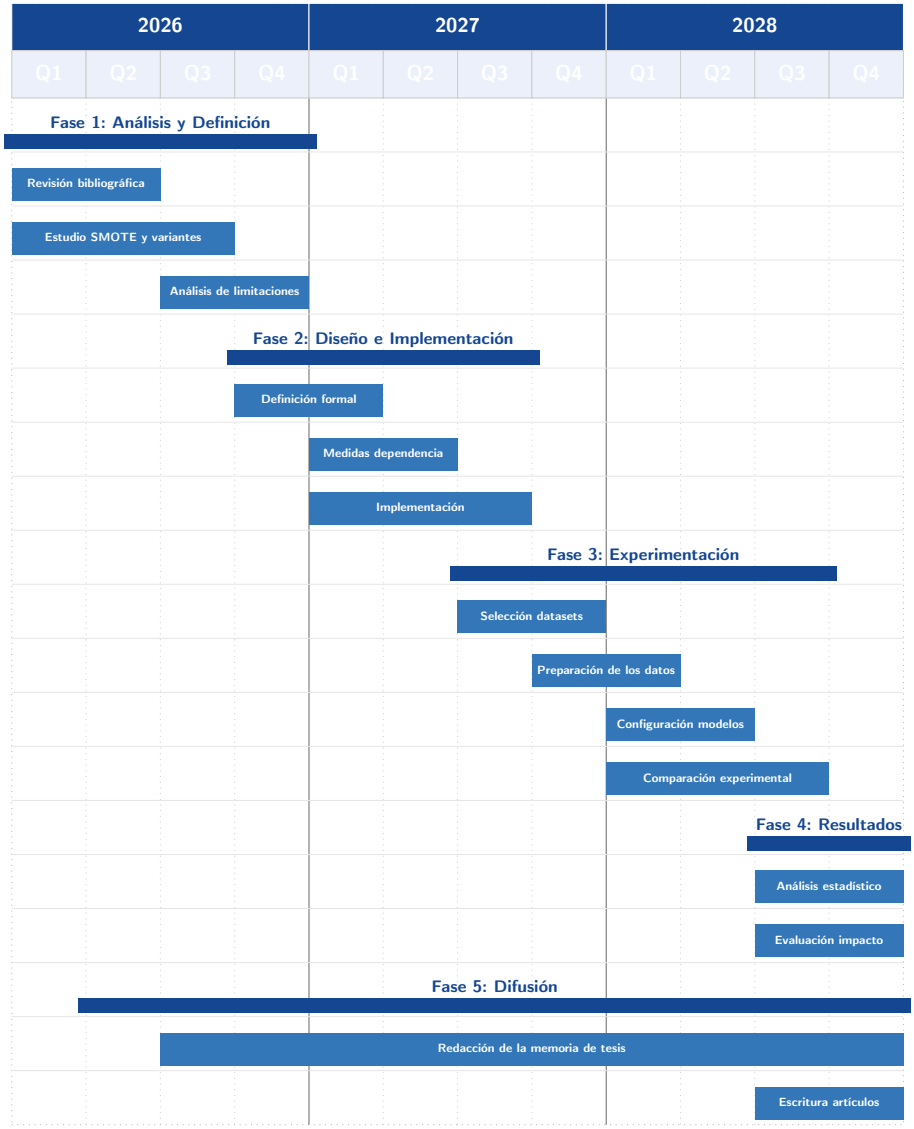


Figura 2. Planificación temporal propuesta.

5. Relevancia

5.1. Relevancia científica y académica

La importancia científica de esta investigación radica en su capacidad para abordar una de las limitaciones más persistentes del aprendizaje automático:

la falta de representatividad en la síntesis de datos para clases minoritarias. A diferencia del estado del arte predominante, centrado principalmente en criterios de proximidad geométrica, esta investigación plantea un cambio de enfoque hacia la generación de muestras basada en la estructura estadística de los datos. Esta propuesta contribuirá al avance del conocimiento al integrar el análisis de dependencias entre varables en el proceso de generación de datos, ofreciendo una perspectiva más rica y fundamentada para el tratamiento del desequilibrio de clases.

Desde una perspectiva académica, el desarrollo de técnicas que reduzcan el ruido sintético y respeten la topología de los datos favorecerá la construcción de modelos más robustos, interpretables y menos sesgados. La integración de información sobre la dependencia entre variables aportará un enfoque novedoso que abrirá nuevas líneas de investigación. Asimismo, los resultados de este trabajo contribuirán a reforzar y ampliar el estado del arte en el tratamiento del desequilibrio de clases dentro del ámbito del aprendizaje automático.

5.2. Transferencia a la sociedad

Más allá del ámbito teórico, esta investigación posee un elevado potencial de transferencia gracias a su aplicabilidad en múltiples dominios y contextos. En una sociedad cada vez más dependiente de la toma de decisiones automatizada, la fiabilidad de los resultados obtenidos mediante la aplicación de la inteligencia artificial es crítica. Al mejorar la detección de eventos poco frecuentes, pero de gran impacto, los resultados de esta investigación podrán integrarse en sistemas de soporte a la decisión en sectores estratégicos.

En el ámbito de la salud, por ejemplo, la capacidad de generar datos sintéticos coherentes facilitará el entrenamiento de modelos para enfermedades raras, donde la escasez de muestras reales suele ser la principal barrera para el avance médico. En el sector de la ciberseguridad y las finanzas, un sobremuestreo más inteligente permitirá reducir significativamente los falsos negativos, protegiendo infraestructuras críticas y ahorrando costes derivados del fraude o los ciberataques. En definitiva, la transferencia de esta tecnología se traducirá en herramientas más seguras y fiables, aumentando la confianza de la sociedad en el despliegue de soluciones basadas en Inteligencia Artificial en escenarios donde el error no es una opción.

Referencias

1. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192
2. Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, Article 106. https://doi.org/10.1186/1471-2105-14-106
3. 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

4. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953
5. Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), *Advances in Intelligent Computing. ICIC 2005* (Lecture Notes in Computer Science, Vol. 3644, pp. 878–887). Springer. https://doi.org/10.1007/11538059_91
6. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 1322–1328). IEEE. https://doi.org/10.1109/IJCNN.2008.4633969
7. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for handling the class imbalanced problem. In T. Theeramunkong, B. Kijsirikul, N. Cercone, & T. B. Ho (Eds.), *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science, Vol. 5476, pp. 475–482). Springer. https://doi.org/10.1007/978-3-642-01307-2_43
8. Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184–203. https://doi.org/10.1016/j.ins.2014.08.051
9. Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503
10. Rodríguez-Torres, F., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2022). An oversampling method for class imbalance problems on large datasets. *Applied Sciences*, *12*(7), Article 3424. https://doi.org/10.3390/app12073424