EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Reentrenamiento de un modelo de lenguaje de gran tamaño como apoyo para la diagnosis del cáncer de tiroides

presentada por

Ing. Alejandro Alvarez Juarez

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Máximo López Sánchez

Codirector de tesis

Dr. Nimrod González Franco

Cuernavaca, Morelos, México. Junio de 2026.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 12/JUNIO/2026
OFICIO No. DCC/123/2026
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **Alejandro Alvarez Juarez** con número de control **M24CE046**,. de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Reentrenamiento de un modelo de lenguaje de gran tamaño como apoyo para la diagnosis del cáncer de tiroides"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

Dr. Máximo López Sánchez
Director de tesis

Dr. Nimrod González Franco
Codirector de tesis

Dr. Jonathan Villanueva Tavira
Revisor 1

Dr. Noé Alejandro Castro Sánchez
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante

2026 año de Margarita Maza

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 15/junio/2026
Oficio No. SAC/163/2026
Asunto: Autorización de impresión de tesis

ALEJANDRO ALVAREZ JUAREZ
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Reentrenamiento de un modelo de lenguaje de gran tamaño como apoyo para la diagnosis del cáncer de tiroides"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico
Interior Internado Palmira S/N, Col. Palmira,
C.P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4106,
e-mail: acad_cenidet@tecnm.mx | cenidet.tecnm.mx

Dedicatoria

A Dios, porque gracias a él todo es posible, bendito sea mi padre celestial que todo lo hace posible, a él dedico mi vida entera, esto ha sido posible porque ha sido su voluntad. Sea hecha su voluntad, como en el cielo, así también en la tierra.

"Encomienda tus obras al Señor y tus propósitos se afianzarán." (Proverbios 16:3)

A mi mamá por su apoyo, respaldo, compresión y todo lo que una madre pueda representar.

Agradecimientos

A Dios por hacer el sueño posible, en todo momento recurrí a el para guiar mi existencia, mis pasos y ayudarme en los momentos difíciles donde parecía tirar la toalla, me ayudaba a reivindicar mi camino.

A Dora mi mamá, por que regrese a casa para impulsarme y ella me apoyó tal como lo hacen las madres, porque con sus cuidados y cariño de madre ha sido posible alcanzar esta meta. Sus consejos y su cocina me motivaban a continuar.

Al Tecnológico Nacional de México (TecNM) por proporcionarme todo lo necesario para mi desarrollo profesional.

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) antes Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico que me brindó para realizar mis estudios de maestría.

Un especial agradecimiento a mi directora de tesis, la Dr. Máximo López Sánchez, por sus consejos, observaciones y paciencia para atender mi trabajo de investigación, además, por su amistad, sentido humano y por haberme enseñado el valor de arriesgarme a lo desconocido.

Un agradecimiento a mi codirector de tesis Dr. Nimrod González Franco por aclarar mis dudas y siempre tener la disposición de ayudarme.

A mis revisores Dr. Noé Alejandro Castro Sánchez y Dr. Jonathan Villanueva Tavira por el tiempo dedicado y contribuir en la revisión de este trabajo, de esta forma lo fortalecieron con sus aportaciones.

A los amigos que hice en el camino y me han apoyado durante el proceso.

A todo el personal administrativo del CENIDET por todo el apoyo y disponibilidad brindado durante mi estancia.

Resumen

El análisis automático de texto clínico en español para el apoyo al diagnóstico oncológico es un área de creciente interés, condicionada por la escasez de recursos lingüísticos especializados en este idioma y por el predominio de herramientas diseñadas principalmente para el inglés. La presente investigación parte de que el conocimiento contenido en los registros clínicos puede aprovecharse mediante modelos de lenguaje biomédicos preentrenados en español, con el propósito de apoyar la clasificación de la posible presencia de cáncer de tiroides a partir de narrativas clínicas.

Se adaptó el modelo de lenguaje biomédico BSC-BIO-EHR-ES mediante un proceso de reentrenamiento sobre un corpus construido a partir de información médica de fuentes abiertas. El desarrollo siguió un protocolo por fases que comprendió la preparación y depuración de los datos, su evaluación mediante algoritmos de clasificación tradicionales, la transformación de los registros estructurados en narrativas clínicas redactadas conforme a la normatividad mexicana aplicable, y el ajuste y la validación del modelo bajo un esquema de validación cruzada anidada.

Los resultados muestran que la representación del conocimiento clínico en lenguaje natural, procesada por el modelo biomédico, aporta mayor valor diagnóstico que el tratamiento tabular con clasificadores clásicos, a los que supera de forma estadísticamente significativa y con un desempeño competitivo dentro de los umbrales establecidos. Se concluye que transformar datos estructurados en narrativas clínicas para activar el conocimiento biomédico preentrenado constituye una estrategia viable y reproducible para el diagnóstico asistido en escenarios de escasez de datos en español.

Palabras clave: procesamiento de lenguaje natural; cáncer de tiroides; modelos de lenguaje biomédicos; clasificación diagnóstica; español clínico.

Abstract

The automatic analysis of clinical text in Spanish to support oncological diagnosis is an area of growing interest, constrained by the scarcity of specialized linguistic resources in this language and by the predominance of tools designed primarily for English. This research is based on the premise that the knowledge contained in clinical records can be leveraged through biomedical language models pretrained in Spanish, with the aim of supporting the classification of the possible presence of thyroid cancer from clinical narratives.

The biomedical language model BSC-BIO-EHR-ES was adapted through a retraining process on a corpus built from medical information drawn from open sources. The development followed a phased protocol comprising the preparation and cleaning of the data, its evaluation using traditional classification algorithms, the transformation of the structured records into clinical narratives written in accordance with the applicable Mexican regulations, and the tuning and validation of the model under a nested cross-validation scheme.

The results show that representing clinical knowledge in natural language, processed by the biomedical model, provides greater diagnostic value than the tabular treatment with classical classifiers, which it outperforms in a statistically significant manner and with a competitive performance within the established thresholds. It is concluded that transforming structured data into clinical narratives in order to activate pretrained biomedical knowledge constitutes a viable and reproducible strategy for assisted diagnosis in low-data scenarios in Spanish.

Keywords: natural language processing; thyroid cancer; biomedical language models; diagnostic classification; clinical Spanish.

ÍNDICE

Pág.

Pág.

Pág.

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

Pág.

GLOSARIO DE TÉRMINOS

A

ANOVA. Analysis of Variance**.**
API. Application Programming Interface.
ARI. Adjusted Rand Index.
ATA. American Thyroid Association.
AUC. Area Under the Curve.
AUROC. Area Under the Receiver Operating Characteristic.

B

BERT. Bidirectional Encoder Representations from Transformers.
BMI. Body Mass Index.
BPE. Byte-Pair Encoding.
BSC. Barcelona Supercomputing Center.
BSC-BIO-EHR-ES. Barcelona Supercomputing Center – Biomedical – Electronic Health Records – Español.

C

CANTEMIST *Cancer Text Mining Shared Task*.
CART. Classification And Regression Trees.
CENIDET. Centro Nacional de Investigación y Desarrollo Tecnológico.
CIE. Clasificación Internacional de Enfermedades.
CLS. Classification Token.
CNN. Convolutional Neural Network.
CoQ. Chain of Questioning.
CoT. Chain of Thought.
CRF. Conditional Random Field.
CSV. Comma-Separated Values.
CUDA. Compute Unified Device Architecture.

D

DEAP. Distributed Evolutionary Algorithms in Python.
DPO. Direct Preference Optimization.

E

ECE. Expected Calibration Error.
EHR. Electronic Health Records.

F

FN. Falso Negativo.
FP. Falso Positivo.

G

GELU. Gaussian Error Linear Unit.
GGUF. GPT-Generated Unified Format.
GPT. Generative Pre-trained Transformer.
GPU. Graphics Processing Unit.

H

HCE. Historia Clínica Electrónica.
HDBSCAN. Hierarchical Density-Based Spatial Clustering of Applications with Noise.

I

IA. Inteligencia Artificial.
IMC. Índice de Masa Corporal.
INEGI. Instituto Nacional de Estadística y Geografía.

K

KL. Kullback-Leibler.
KNN. K-Nearest Neighbors.

L

LLM. Large Language Model.
LoRA. Low-Rank Adaptation.
LR. Learning Rate.

M

mBERT. Multilingual BERT.
MCC. Matthews Correlation Coefficient.
MDPI. Multidisciplinary Digital Publishing Institute.
MLM. Masked Language Modeling.
MLP. Multilayer Perceptron.
MMD. Maximum Mean Discrepancy.

N

NCI. National Cancer Institute.
NER. Named Entity Recognition.
NIH. National Institutes of Health.
NLP. Natural Language Processing.
NOM. Norma Oficial Mexicana.
NSP. Next Sentence Prediction.

O

OMS. Organización Mundial de la Salud.
OPS. Organización Panamericana de la Salud.

P

PCA. Principal Component Analysis.
PLN. Procesamiento de Lenguaje Natural.
PR-AUC. Precision-Recall Area Under the Curve.

R

RAG. Retrieval-Augmented Generation.
RAM. Random Access Memory.
ReAct. Reasoning and Acting.
ReLU. Rectified Linear Unit.
RoBERTa. Robustly Optimized BERT Approach.
ROC. Receiver Operating Characteristic.
ROC-AUC. Receiver Operating Characteristic – Area Under the Curve.

S

SEER. Surveillance, Epidemiology, and End Results.
SFT. Supervised Fine-Tuning.
SMB. Server Message Block.
SMOTE. Synthetic Minority Over-sampling Technique.
SMTP. Sequential Masked Token Prediction.
SOTA. State Of The Art.
SSH. Secure Shell.
SVM. Support Vector Machine.

T

T3. Triyodotironina.
T4. Tiroxina.
TCGA The Cancer Genome Atlas**.**
THCA Thyroid Carcinoma.
TN. True Negative.
TP. True Positive.

U

UAEMex. Universidad Autónoma del Estado de México.
UFW. Uncomplicated Firewall.

V

vCPU. Virtual Central Processing Unit.
VN. Verdadero Negativo.
VP. Verdadero Positivo.
VPS. Virtual Private Server.

X

XLM-R. Cross-lingual Language Model – RoBERTa.
XML. Extensible Markup Language.

CAPÍTULO I
INTRODUCCIÓN

El cáncer es un padecimiento de salud que ha aquejado a la humanidad, causando estragos que muchas veces se consideran irreversibles a nivel mundial. De acuerdo con el *INEGI* en México durante el 2022 se registraron 89,579 defunciones por tumores malignos, lo que representa el 10.6% del total de muertes en el país. La tendencia va en incremento dado que la mortalidad por cáncer pasó de 62 muertes por cada 100,000 habitantes en 2012 a 68 en 2022. (I.N.E.G.I., 2024) En ese contexto, cada mejora en detección temprana, cada minuto ahorrado en el diagnóstico y cada decisión clínica mejor informada puede significar una diferencia real. Dentro de este panorama, el cáncer de tiroides ocupa un lugar particular no es el más frecuente, pero sí exige una precisión quirúrgica en su abordaje. Su diagnóstico y tratamiento pueden ser complejos, y sus efectos impactan de forma directa en la calidad de vida. Se caracteriza por el crecimiento descontrolado de células anormales en la glándula tiroides, y su pronóstico depende, en gran medida, de qué tan rápido y qué tan bien se identifique. Entre sus factores de riesgo se incluyen la predisposición genética, la exposición a radiaciones ionizantes y los desequilibrios hormonales (Hundahl et al., 1998). En México, el manejo de esta información clínica no solo es médico: también es normativo y ético, pues se rige por lineamientos como la *NOM-004-SSA3-2012*(Secretaría de Salud, 2012) que establece directrices para el expediente clínico electrónico y el tratamiento de información médica. Con base a lo anterior, la información al encontrarse en formato de texto y en lenguaje natural, puede ser procesada por un modelo de lenguaje de gran tamaño (*LLM* por sus siglas en inglés) una herramienta capaz de leer e interpretar información en dicho formato (Vaswani et al., 2017). En años recientes, modelos especializados en temas médicos y clínicos como lo son las variantes de *Bidirectional Encoder Representations from Transformers* (*BERT*) (Devlin et al., 2019) han mostrado alta precisión al detectar señales oncológicas a partir de notas clínicas electrónicas (Alsentzer et al., 2019). Sin embargo, su principal limitación es tan simple como determinante: fueron entrenados exclusivamente en inglés (J. Lee et al., 2020). Eso restringe su utilidad en contextos hispanohablantes, justo donde la necesidad de herramientas accesibles y contextualizadas puede ser mayor.

La barrera lingüística no es un detalle menor; es un obstáculo técnico de fondo. Los *tokenizadores* optimizados para inglés fragmentan de forma ineficiente el vocabulario médico en español (Agerri et al., 2023), debilitando la capacidad del modelo para capturar el sentido clínico de los términos. Además, la práctica médica en Latinoamérica tiene su propia forma de escribirse: abreviaturas, giros, sintaxis y convenciones que no aparecen en reportes anglosajones (Miranda-Escalada et al., 2022). En otras palabras: no basta con "traducir" el problema; hay que entrenar soluciones que entiendan el idioma y el contexto en el que viven los pacientes y trabajan los médicos. Bajo esta premisa, el presente trabajo tiene como objetivo realizar el ajuste fino (*fine-tuning*) de un modelo de lenguaje grande especializado para el análisis de información médica relacionada con el cáncer de tiroides en español, con enfoque en el contexto mexicano. En particular, se propone adaptar el modelo *BSC-BIO-EHR-ES* (Carrino et al., 2022), preentrenado con registros clínicos electrónicos en español, para una tarea de clasificación diagnóstica. Con ello, se busca facilitar el procesamiento de historiales clínicos no estructurados e identificar patrones que apoyen el diagnóstico temprano, donde la oportunidad clínica suele ser decisiva.

La presente tesis tiene como propósito resaltar una intersección entre inteligencia artificial, análisis de información médica y la importancia de la representación del lenguaje natural, en donde es posible visualizar el potencial de la inteligencia artificial expresada en un modelo de lenguaje como una posible herramienta viable e implementable en un entorno clínico.

1.1 Planteamiento del problema

En México la detección del cáncer de tiroides en una etapa temprana es fundamental. Desafortunadamente las herramientas computacionales basadas en el procesamiento de lenguaje, se han especializado sobre el contexto del idioma inglés, lo que limita su aplicación bajo las condiciones de otro idioma como el español (Alsentzer et al., 2019), la falta de modelos entrenados con datos locales y en el idioma antes mencionado limita su implementación efectiva en sistemas de salud regionales.

Ante esta situación, es urgente desarrollar soluciones tecnológicas que permitan:

- **Procesar y analizar datos clínicos no estructurados en español**, extrayendo información clave sobre síntomas, diagnósticos y tratamientos.

- **Adaptar y entrenar modelos de lenguaje existentes con datos clínicos mexicanos**, incluyendo registros médicos electrónicos y reportes patológicos, para reflejar las particularidades epidemiológicas locales.
- **Mejorar la toma de decisiones clínicas mediante herramientas que proporcionen información personalizada y actualizada**, basada en datos específicos de cada paciente y en el conocimiento médico vigente.

1.2 Justificación

El cáncer es una de las principales causas de mortalidad en las Américas. En el 2022, causó 1,4 millones de muertes, un 45,1% de ellas en personas de 69 años de edad o más jóvenes (Organización Panamericana de la Salud, 2024) El número de casos de cáncer en la Región de las Américas se estimó en 4,2 millones en 2022 y se proyecta que aumentará hasta los 6,7 millones en 2045(Organización Panamericana de la Salud, 2024).

Datos/Estadísticas
40% de casos podrían prevenirse evitando factores de riesgo clave
30% de casos pueden curarse si se detectan temprano y se tratan adecuadamente (Organización Panamericana de la Salud, 2024).

Quienes reciben un diagnóstico de cáncer en estado temprano tienen una probabilidad del 91% de seguir con vida a los 5 años del diagnóstico en comparación con quienes no tienen este tipo de cáncer (Organización Panamericana de la Salud, 2024).

De acuerdo al *INEGI* en 2022, en México se registraron 847 716 defunciones 10.6 % fue por tumores malignos (89 574) como es posible apreciar en la Figura 1.1. La tasa de defunciones por esta causa aumentó de forma constante, al pasar de 62.04 defunciones por cada 100 mil personas en 2012, a 68.92 en 2022(I.N.E.G.I., 2024).

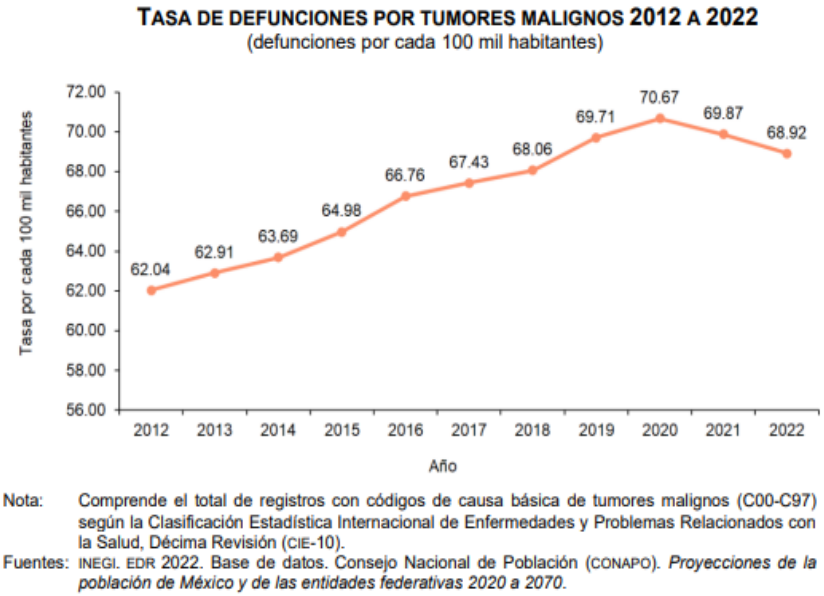

Figura 1.1: Grafico donde se muestra la tasa de defunción por tumores malignos del 2012 a 2022

1.3 Objetivos

1.3.1 Objetivo general

Adaptar un modelo *LLM* como apoyo a la clasificación de la posible presencia de cáncer de tiroides mediante su reentrenamiento a partir de bases de datos e información médica, para aportar información diagnóstica y completar la validación del modelo.

1.3.2 Objetivos específicos

- (OE1) Observar la normatividad ética y legal vigente para compilar y gestionar corpus de textos en español relacionados con los casos de cáncer en México, incluyendo diagnóstico, tratamiento y resultados clínicos.
- (OE2) Detectar patrones y particularidades que puedan afectar la generación de información del modelo *LLM*.
- (OE3) Optimizar la capacidad de generar información sobre el diagnóstico del cáncer en español.
- (OE4) Comparar con métodos de referencia.
- (OE5) Identificar fortalezas, limitaciones y áreas de mejora en el modelo y su aplicación en el mundo real.

1.4 Alcances y Limitaciones

1.4.1 Alcances

El desarrollo de esta tesis tiene como objetivo identificar las posibles características de la sintomatología del cáncer de tiroides mediante el reentrenamiento de modelos *LLM*. Los alcances del trabajo abarcan tanto los aspectos técnicos como las aplicaciones prácticas de la solución propuesta.

1.4.1.1 Aproximación al apoyo en la Detección y Diagnóstico del Cáncer

- Al entrenar el modelo con un corpus clínico que incluye datos relevantes sobre síntomas, diagnósticos y tratamientos del cáncer, se espera que el modelo proporcione una herramienta útil para procesar grandes volúmenes de datos clínicos no estructurados. Esto incluye la capacidad de identificar patrones clínicos en registros médicos electrónicos (*EHR*) (Jensen et al., 2012), facilitando el diagnóstico temprano y preciso.
- Este alcance contribuye a reducir el tiempo necesario para analizar datos clínicos, mejorando la eficiencia en la toma de decisiones médicas.

1.4.1.2 Desarrollo de una Herramienta Adaptada al Idioma Español

- Gran parte de los modelos de lenguaje existentes, como *BERT*(Devlin et al., 2019) y *GPT*(Brown et al., 2020), están diseñados para trabajar en inglés, lo que los hace menos útiles en contextos hispanohablantes. El reentrenamiento en español utilizando datos médicos locales permitirá al modelo comprender términos médicos específicos y sus variaciones lingüísticas.
- Esto asegura una mayor aplicabilidad del modelo en hospitales y centros médicos mexicanos, donde el español es el idioma predominante.

1.4.1.3 Personalización para el Contexto Mexicano

- Los tipos de cáncer más prevalentes en México, como el de tiroides, el de próstata y el de mama, pueden diferir de otras regiones. Adaptar el modelo con datos locales permitirá reflejar particularidades epidemiológicas, culturales y sociales del país.
- Al abordar las necesidades específicas del sistema de salud mexicano, se espera que el modelo sea clínicamente relevante y alineado con los recursos y realidades locales.

1.4.1.4 Prototipo Funcional

- Este prototipo facilitará la integración del modelo en los flujos de trabajo clínicos y servirá como una prueba de concepto para posibles implementaciones a gran escala.

1.4.2 Limitaciones

A pesar de los avances esperados, este trabajo enfrenta varias limitaciones inherentes a los datos, la tecnología y el contexto en el que se desarrollará. Estas limitaciones son importantes para establecer expectativas realistas y enfocar los esfuerzos de la investigación.

1.4.2.1 Disponibilidad de Datos Clínicos en español

- Existe una escasez de bases de datos abiertas y extensas en español relacionadas con casos de cáncer (Névéol et al., 2018). Además, la obtención de registros médicos en México puede estar restringida por normativas legales y de privacidad.
- La falta de datos suficientes podría limitar el tamaño del corpus *Gold Standard* y afectar la capacidad del modelo para generalizar en escenarios diversos.
- Se mitigará esta limitación mediante la generación de datos sintéticos (Gonzalez-Agirre et al., 2019) que complementen los datos reales recopilados.

1.4.2.2 Calidad de las Anotaciones

- La anotación manual de datos clínicos es un proceso que consume tiempo y depende de la experiencia de los anotadores. Esto puede llevar a inconsistencias o errores en el corpus *Gold Standard*.
- Errores en la anotación podrían reducir la precisión del modelo y su capacidad para identificar patrones relevantes en los datos.
- Se realizarán revisiones exhaustivas y validaciones cruzadas para garantizar la calidad y consistencia del corpus.

1.4.2.3 Requisitos Computacionales

- Entrenar y ajustar un modelo de lenguaje grande requiere recursos computacionales significativos, como *GPUs* de alto rendimiento (Strubell et al., 2019), que pueden no estar completamente disponibles.
- Limitaciones en los recursos computacionales podrían alargar los tiempos de entrenamiento o requerir simplificaciones en el diseño del modelo.
- Se utilizarán plataformas en la nube y optimizaciones de modelos para maximizar la eficiencia computacional.

1.4.2.4 Generalización del Modelo

- Aunque el modelo será *re-entrenado* con datos específicos del contexto mexicano, podría enfrentar dificultades para generalizar a otros contextos clínicos o países donde las características epidemiológicas o la terminología médica difieren.
- Esto limita el uso del modelo fuera de México o en dominios médicos distintos al cáncer.
- El objetivo principal de la investigación es atender las necesidades locales, por lo que esta limitación es aceptable en el marco del proyecto.

1.4.2.5 Complejidad de la Interpretación Médica

- Las decisiones clínicas no solo dependen de datos estructurados, sino también de juicios contextuales, experiencia médica y consideraciones éticas que un modelo de lenguaje no puede reemplazar (Topol, 2019).
- El modelo no puede actuar como una herramienta de diagnóstico autónoma y requiere supervisión médica constante.
- El modelo está diseñado como un sistema de apoyo a la decisión clínica, no como un sustituto del juicio médico.

1.5 Antecedentes

Anteriormente en el Centro Nacional de Investigación y Desarrollo Tecnológico (*CENIDET*), se desarrollaron trabajos de tesis enfocados en la detección de enfermedades a través de alteraciones observadas

en el iris del ojo. A continuación, se describen dos investigaciones relevantes que sirvieron como antecedentes para este trabajo:

En **Pineda Tapia (2016)** (J.A.Pineda Tapia, 2016) se presenta el desarrollo de una aplicación bajo la plataforma Android para la detección de **cáncer abdominal** mediante el análisis de alteraciones en el iris. El proceso constó de cuatro fases principales:

- **Captura de la imagen y detección del iris** Se utilizaron algoritmos como *Canny* para detectar bordes y la transformada de *Hough* para identificar estructuras circulares como el iris.
- **Preprocesamiento y procesamiento** Eliminación de áreas no relevantes, segmentación de la zona pupilar e irídica, y extracción de características para identificar alteraciones. El iris segmentado se dividió en cuatro cuadrantes, enfocándose en el tercer cuadrante donde se manifiestan las alteraciones asociadas al cáncer abdominal.
- **Consulta y clasificación** Se utilizó una base de datos de 5,008 imágenes, clasificadas en cuatro etapas de evolución del cáncer abdominal. Se entrenaron varios clasificadores para identificar las alteraciones con base en estas imágenes.
- **Pruebas y validación** Se alcanzó un porcentaje de asertividad del 98% en la detección de alteraciones relacionadas con el cáncer abdominal, mostrando el potencial del iris como biomarcador no invasivo.

Por otro lado, en **Rosales Banderas (2017)** (Banderas, 2017) se desarrolló una aplicación para detectar la presencia del anillo de colesterol en el iris, un signo de hiperlipidemia. El trabajo incluyó dos etapas:

- **Captura y detección del iris** Se utilizaron smartphones como dispositivos de captura, aplicando algoritmos de procesamiento de imágenes para identificar bordes y circunferencias.
- **Análisis del iris segmentado** Se empleó el modelo de color *HSV* (Smith, 1978) para medir la coloración y saturación en el área del anillo. Se logró un 81.08% de precisión en la detección del anillo de colesterol, demostrando la capacidad de los dispositivos móviles y técnicas de procesamiento para tareas biomédicas.

Estos trabajos resaltan el potencial del iris como fuente de información para el diagnóstico médico y la integración de tecnologías accesibles como los smartphones y algoritmos de inteligencia artificial para abordar problemas clínicos de manera no invasiva.

CAPÍTULO II:
MARCO TEÓRICO

En este capítulo se exponen los fundamentos teóricos y conceptuales sobre los que se apoya la investigación. Así como son abordados los conceptos clínicos del cáncer de tiroides y la normatividad mexicana aplicable al mismo. Se exponen, a continuación, las bases del Procesamiento del Lenguaje Natural, las técnicas de representación textual mediante "*embeddings*" y los paradigmas de aprendizaje automático, incluyendo los algoritmos de agrupamiento y clasificación empleados como línea base. En la sección central se detalla la arquitectura *Transformer* y los modelos *BERT*, *RoBERTa* y *BSC-BIO-EHR-ES*, así como las estrategias de transferencia de conocimiento y ajuste fino. Las técnicas de reducción de dimensionalidad (*PCA* y *t-SNE*) son expuestas, así como las métricas de evaluación, incluyendo las pruebas estadísticas no paramétricas de *Wilcoxon* y *Friedman* para validar los resultados obtenidos.

2.1 Dominio clínico y normativo

2.1.1 Cáncer de tiroides: contexto clínico

La glándula tiroides es una glándula endocrina localizada en la cara anterior del cuello cuya principal función es la producción de hormonas tiroideas (*triyodotironina*, *T3*, y *tiroxina*, *T4*) que regulan funciones metabólicas esenciales del organismo. El cáncer de tiroides es el crecimiento excesivo de células anormales en esta glándula, y su incidencia ha ido en aumento sostenido en las últimas décadas, sobre todo en Latinoamérica.

Desde el punto de vista histopatológico, los tumores malignos de tiroides se clasifican en cuatro tipos principales: *papilar* (80% de los casos), *folicular*, *medular* y *anaplásico*. Los factores de riesgo más importantes son la predisposición genética, la exposición a radiaciones ionizantes, los desequilibrios hormonales y los antecedentes familiares de enfermedad de la glándula tiroides. El Instituto Nacional de Estadística y Geografía (*INEGI*) informó que durante 2022 en México se registraron 89 mil 579 muertes por tumores malignos, es decir, el 10.6% del total de los fallecimientos en el país. La mortalidad por esta causa pasó de ser 62 muertes por cada 100.000 habitantes en 2012, a ser 68 en 2022 (I.N.E.G.I., 2024), lo que resalta la importancia de diseñar herramientas tecnológicas que ayuden al diagnóstico temprano y al apoyo diagnóstico.

2.1.2 Normatividad: NOM-004-SSA3-2012

En México, el manejo de la información clínica está regulado por la Norma Oficial Mexicana *NOM-004-SSA3-2012* (Secretaría de Salud, 2012), que establece las directrices para la integración, uso, archivo y confidencialidad del expediente clínico. En el contexto de este trabajo de investigación, la norma tiene una doble importancia: por un lado, establece el marco legal bajo el cual se creó y se manejó el corpus de textos clínicos y, por otro lado, determina la estructura y el estilo de redacción de las notas clínicas mexicanas, que difieren sustancialmente de los reportes médicos anglosajones en su sintaxis, uso de abreviaturas y convenciones terminológicas. Estas características lingüísticas son uno de los argumentos principales para justificar la necesidad de un modelo de lenguaje entrenado específicamente con registros clínicos en español.

2.2 Procesamiento de lenguaje natural y aprendizaje automático

2.2.1 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (*PLN*), conocido en inglés como *Natural Language Processing* (*NLP*), es una subdisciplina de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano (Jurafsky & Martin, 2026). Su objetivo fundamental consiste en dotar a las máquinas de la capacidad de comprender, interpretar y generar texto de manera significativa. El *PLN* abarca tareas como la clasificación de textos, el reconocimiento de entidades nombradas (*NER*), el análisis de sentimientos, la traducción automática y la respuesta a preguntas (Jurafsky & Martin, 2026). En el dominio biomédico, el *PLN* se ha convertido en una herramienta indispensable para la extracción de información a partir de registros clínicos electrónicos y notas médicas no estructuradas, dado que una proporción significativa de la información médica se almacena en formato de texto libre.

En este contexto, la transformación de los textos libres a formatos estructurados se logra fundamentalmente mediante la extracción de información, una tarea que permite descubrir y clasificar relaciones semánticas complejas dentro del texto. Por ejemplo, en el entorno médico, los sistemas de PLN pueden extraer relaciones directamente de oraciones clínicas para determinar automáticamente que un procedimiento como la "ecocardiografía Doppler" diagnostica una "estenosis". Esta capacidad es un paso esencial para poblar bases de datos relacionales y construir grafos de conocimiento a partir de historiales de pacientes y literatura especializada (Jurafsky & Martin, 2026).

A la par de la extracción de entidades y relaciones, la clasificación de textos permite categorizar los documentos de forma automática mediante algoritmos de aprendizaje automático supervisado. Dentro de estas tareas destaca el análisis de sentimientos, cuyo propósito es extraer la orientación afectiva, las opiniones o las evaluaciones de un autor. Aunque es muy popular en reseñas de productos o foros políticos, la clasificación textual también abarca dominios como la detección de spam, la identificación de idiomas y la atribución de autoría (Jurafsky & Martin, 2026).

Para facilitar el análisis y consumo global de toda esta información, la traducción automática actúa como una herramienta fundamental que ayuda a reducir la brecha digital. En la actualidad, estos sistemas traducen cientos de miles de millones de palabras al día, permitiendo cruzar barreras idiomáticas para acceder a instrucciones, artículos de noticias y registros técnicos en una vasta cantidad de lenguas (Jurafsky & Martin, 2026).

Por su parte, el campo de la respuesta a preguntas se ha visto transformado por la integración de los Modelos de Lenguaje Grandes (LLMs), los cuales aprenden el conocimiento del mundo preentrenándose para predecir palabras a partir de contextos masivos. Sin embargo, en ámbitos como la salud o el derecho la exactitud de los datos es crítica, y los LLMs son propensos a "alucinar" (afirmar hechos falsos), lo que puede generar riesgos significativos. Para mitigar este problema, el PLN moderno emplea técnicas como la Generación Aumentada por Recuperación (RAG). La arquitectura RAG resuelve las limitaciones de los modelos estáticos al integrar técnicas de recuperación de información que buscan en fuentes propietarias y fidedignas (como notas médicas, correos personales o bases de conocimiento empresariales) para fundamentar las respuestas del modelo en evidencias textuales reales (Jurafsky & Martin, 2026).

En definitiva, integrando desde la tokenización y el análisis morfológico base hasta las arquitecturas neuronales bidireccionales y predictivas como los Transformers, el PLN logra construir representaciones vectoriales profundas sobre el significado de las palabras en su contexto. Esto consolida a la disciplina como un eje tecnológico indispensable para automatizar la comprensión y generación del lenguaje de forma segura, útil y significativa (Jurafsky & Martin, 2026).

En el siguiente diagrama contenido en la Figura 2.1 representa todo el procesamiento del lenguaje natural, describiendo las etapas anteriormente mencionadas.

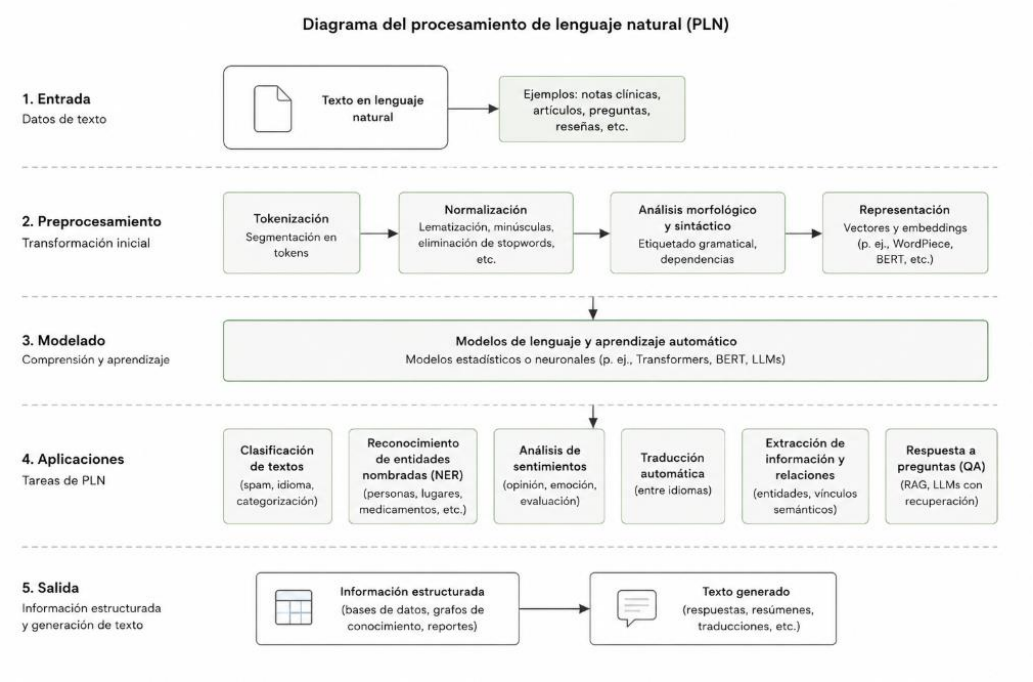

Figura 2.1: Diagrama con la representación de todo el procesamiento que se realiza del lenguaje natural, comenzado por la entrada de información y culminando en un producto como salida.

2.2.2 Aprendizaje automático

El aprendizaje automático (*Machine Learning*) es una rama de la inteligencia artificial que permite a los algoritmos mejorar su rendimiento a partir de la experiencia, sin ser explícitamente programados para ello. En función de la naturaleza de los datos y la retroalimentación proporcionada, se clasifica en tres paradigmas: supervisado, no supervisado y por refuerzo (Bishop, 2006). En esta investigación se emplearon los dos primeros como línea base antes de implementar modelos de aprendizaje profundo. El aprendizaje supervisado se caracteriza por el uso de conjuntos de datos previamente etiquetados, en los cuales cada instancia de entrada se encuentra asociada con una salida esperada. A partir de esta relación, el modelo aprende una función capaz de generalizar patrones y realizar predicciones sobre nuevos datos no observados durante el entrenamiento. Dentro de este paradigma se incluyen tareas como la clasificación y la regresión, las cuales resultan fundamentales en problemas donde se busca asignar una categoría específica o estimar un valor continuo a partir de determinadas variables de entrada (Bishop, 2006). En el contexto de esta investigación, este enfoque permitió establecer una primera aproximación al comportamiento predictivo de los datos, evaluando la capacidad de distintos algoritmos para identificar regularidades relevantes.

Por otra parte, el aprendizaje no supervisado se aplica cuando los datos no cuentan con etiquetas predefinidas, por lo que el objetivo principal consiste en descubrir estructuras internas, agrupamientos o relaciones latentes dentro del conjunto de datos. Este paradigma incluye técnicas como el agrupamiento, la estimación de densidad y la reducción de dimensionalidad, las cuales permiten explorar la distribución de la información sin depender de una variable objetivo previamente definida (Bishop, 2006). Su utilidad radica en que facilita la identificación de patrones ocultos, similitudes entre observaciones y posibles subgrupos dentro de los datos, lo cual puede servir como una etapa exploratoria antes de la construcción de modelos predictivos más complejos.

Desde la perspectiva planteada por Bishop (2006), el aprendizaje automático no sólo se limita a la aplicación mecánica de algoritmos, sino que implica la construcción de modelos capaces de representar la

incertidumbre, estimar parámetros y evaluar la capacidad de generalización. En este sentido, resulta necesario considerar aspectos como el ajuste del modelo, la selección de variables, el riesgo de sobreajuste y la validación del desempeño mediante datos no utilizados durante el entrenamiento. Estos elementos son esenciales para evitar que el modelo memorice patrones particulares del conjunto de entrenamiento y, en cambio, logre capturar relaciones consistentes que puedan mantenerse en nuevos casos.

En consecuencia, el uso de modelos supervisados y no supervisados como línea base permitió establecer un punto de comparación inicial frente a arquitecturas de aprendizaje profundo. Esta estrategia resulta pertinente porque facilita evaluar si los modelos más complejos aportan una mejora real en el desempeño o si, por el contrario, los patrones presentes en los datos pueden ser capturados adecuadamente mediante métodos tradicionales de aprendizaje automático. Así, la incorporación posterior de modelos profundos no se plantea como una sustitución automática de los enfoques clásicos, sino como una etapa metodológica orientada a contrastar su capacidad representacional, su rendimiento predictivo y su utilidad dentro del problema de investigación.

2.2.2.1 Algoritmos de agrupamiento no supervisado

El aprendizaje no supervisado busca descubrir estructuras ocultas en los datos sin utilizar etiquetas (Bishop, 2006). En esta investigación se evaluaron dos algoritmos de agrupamiento para determinar si los patrones clínicos presentaban una separabilidad geométrica o de densidad natural.

El primer enfoque permitió analizar la posible formación de grupos a partir de la proximidad entre observaciones dentro del espacio de características. Este tipo de método parte del supuesto de que los datos pueden organizarse en regiones relativamente compactas, donde los elementos de un mismo grupo presentan mayor similitud entre sí que con respecto a los elementos de otros grupos. Bajo esta lógica, el agrupamiento funciona como una herramienta exploratoria para identificar si existen patrones clínicos diferenciables sin necesidad de utilizar una variable objetivo previamente definida.

El segundo enfoque se orientó a evaluar la existencia de agrupamientos determinados por zonas de mayor concentración de datos. A diferencia de los métodos basados únicamente en distancia o centroides, este tipo de aproximación permite reconocer estructuras con formas menos regulares y detectar observaciones que no se integran claramente en ningún grupo. Esto resulta especialmente relevante en datos clínicos, donde los patrones no siempre siguen distribuciones lineales o geométricamente simples, y donde la presencia de casos atípicos puede aportar información importante sobre la heterogeneidad del conjunto analizado.

K-Means

K-Means es un algoritmo de particionamiento iterativo que divide el conjunto de datos en K grupos disjuntos, minimizando la suma de distancias cuadráticas entre cada punto y el centroide de su grupo asignado. Asume clústeres de forma esférica y varianza similar (Bishop, 2006). En esta investigación se configuró con K=2, correspondientes a las dos clases diagnósticas.

El funcionamiento de K-Means se basa en la asignación iterativa de cada observación al centroide más cercano y en la posterior actualización de dichos centroides a partir del promedio de los puntos pertenecientes a cada grupo. Este proceso se repite hasta alcanzar un criterio de convergencia, generalmente cuando las asignaciones dejan de cambiar de forma significativa o cuando la reducción de la función objetivo se vuelve mínima. De esta manera, el algoritmo busca encontrar una partición interna de los datos que reduzca la variabilidad dentro de cada grupo y aumente la separación relativa entre ellos.

En el contexto de esta investigación, la elección de K=2 respondió a la existencia de dos clases diagnósticas previamente definidas, con el propósito de explorar si los datos clínicos presentaban una estructura geométrica compatible con dicha división. Aunque K-Means no utiliza las etiquetas durante el proceso de agrupamiento, la comparación posterior entre los grupos generados y las clases diagnósticas permitió valorar si existía una correspondencia aproximada entre la organización interna de los datos y la clasificación clínica esperada.

No obstante, este algoritmo presenta limitaciones importantes. Su desempeño depende de la inicialización de los centroides, de la escala de las variables y de la forma de los grupos presentes en el espacio de características. Debido a que K-Means favorece agrupamientos compactos, convexos y aproximadamente esféricos, puede presentar dificultades cuando los datos poseen distribuciones irregulares, solapamiento entre clases, diferencias marcadas de densidad o presencia de valores atípicos. Por esta razón, sus resultados deben interpretarse como una aproximación exploratoria y no como una confirmación definitiva de separabilidad clínica. En la Figura 2.2 es posible percibir el funcionamiento del algoritmo K-Means por etapas.

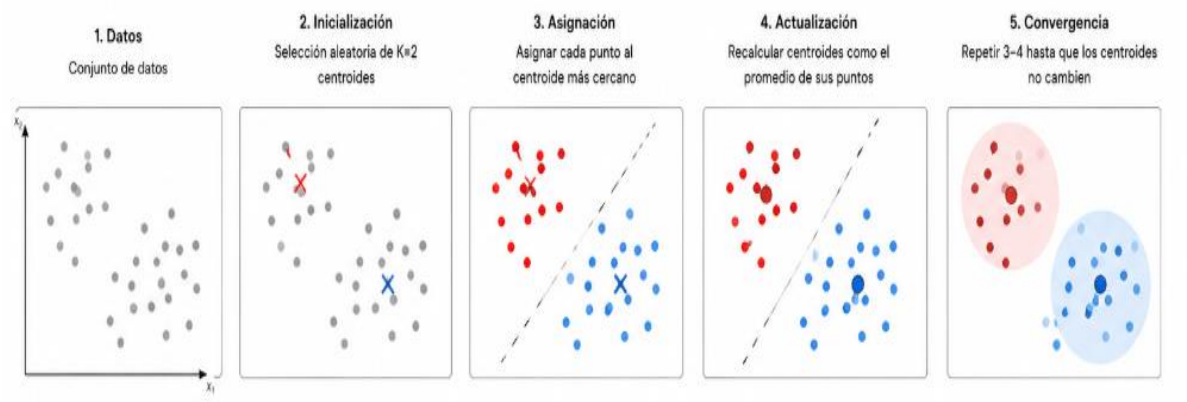

Figura 2.2: Representación de funcionamiento del algoritmo de particionamiento usando K-Means

HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) es un algoritmo jerárquico basado en densidad que, a diferencia de K-Means, no asume formas predefinidas de los clústeres. Construye una jerarquía basada en la densidad de puntos, lo que le permite identificar estructuras de forma arbitraria y descartar valores atípicos como ruido (Bishop, 2006). Su capacidad para detectar clústeres de densidad variable lo convierte en un complemento natural a K-Means.

El principio de HDBSCAN parte de la idea de que los grupos pueden entenderse como regiones del espacio de datos donde existe una mayor concentración de observaciones, separadas por zonas de menor densidad. A partir de esta lógica, el algoritmo no requiere definir previamente el número de clústeres, sino que identifica las agrupaciones más estables dentro de una estructura jerárquica. Esta característica resulta especialmente útil cuando los datos no presentan una separación lineal, cuando los grupos tienen formas irregulares o cuando existen observaciones que no pertenecen claramente a ningún agrupamiento.

En el contexto de esta investigación, HDBSCAN se utilizó como una alternativa más flexible frente a K-Means, debido a que los datos clínicos pueden presentar distribuciones heterogéneas, solapamientos entre clases y patrones de densidad no uniformes. Mientras que K-Means tiende a favorecer clústeres compactos y aproximadamente esféricos, HDBSCAN permite explorar si existen agrupamientos más complejos, definidos por la concentración local de los datos y no únicamente por la distancia respecto a un centroide. En la Figura 2.3 es posible percibir el funcionamiento del algoritmo HDBSCAN por etapas.

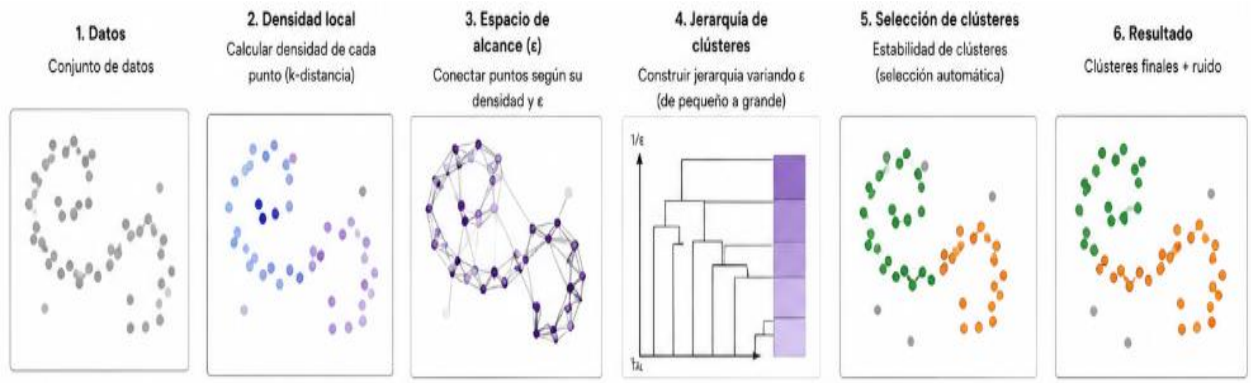

Figura 2.3: Grafico donde se muestra la tasa de defunción por tumores malignos del 2012 a 2022Representación de funcionamiento del algoritmo de particionamiento usando HDBSCAN

2.2.2.2 Algoritmos supervisados de clasificación

En el aprendizaje supervisado, el modelo recibe ejemplos etiquetados y aprende una función de mapeo para predecir la salida de nuevas entradas (Bishop, 2006). La clasificación binaria, caso particular de esta investigación, consiste en asignar cada historial clínico a una de dos clases: presencia o ausencia de cáncer de tiroides. La Tabla 2.1 presenta los seis algoritmos evaluados como línea base supervisada.

Este paradigma parte de un conjunto de entrenamiento compuesto por observaciones de entrada y sus respectivas salidas conocidas. A partir de estos ejemplos, el modelo ajusta sus parámetros con el propósito de capturar regularidades presentes en los datos y utilizarlas posteriormente para inferir la clase correspondiente de nuevos casos. En este sentido, el aprendizaje supervisado no se limita a memorizar los ejemplos disponibles, sino que busca construir una representación capaz de generalizar hacia datos no observados durante el entrenamiento (Bishop, 2006).

En los problemas de clasificación, la salida esperada corresponde a una categoría discreta. Cuando existen únicamente dos posibles clases, el problema se define como clasificación binaria. En esta investigación, dicha formulación permitió representar el diagnóstico clínico como una tarea de decisión entre dos estados: presencia o ausencia de cáncer de tiroides. Esta simplificación metodológica facilitó la comparación entre modelos, ya que todos los algoritmos evaluados fueron entrenados bajo el mismo objetivo predictivo y sobre la misma estructura de datos.

Tabla 2.1: Algoritmos clásicos de clasificación evaluados como línea base supervisada.

Algoritmo	Fundamento	Característica Clave
Regresión Logística	Función sigmoide para estimar probabilidad de clase	Estándar en probabilidad clínica; separación lineal
Árbol de Decisión	Particiona el espacio con reglas jerárquicas si-entonces	Captura relaciones no lineales entre variables
Random Forest	Ensemble: combina múltiples árboles con *Bagging*	Reduce varianza promediando 100 estimadores
SVM (Kernel RBF)	Hiperplano óptimo; *Kernel Trick* para no linealidad	Proyecta a dimensiones superiores para separación
KNN (k=5)	Clase mayoritaria entre los k vecinos más cercanos	Aprendizaje perezoso; depende de geometría local
Naive Bayes	Teorema de Bayes con independencia condicional	Eficiente, pero con supuesto simplificador fuerte

Como se observa en la Tabla 2.1, cada algoritmo opera bajo supuestos diferentes sobre la naturaleza de los datos. Los resultados experimentales, presentados en el Capítulo 4, demostrarán que ninguno superó la barrera del azar (~50%) sobre los datos tabulares del *Dataset* de 534 muestras, evidenciando que la información diagnóstica no reside en la topología vectorial de los datos tabulares sino en su semántica, lo que justificó la transición hacia modelos de lenguaje profundos.

Los algoritmos supervisados evaluados como línea base representan distintas formas de aproximar la relación entre las variables de entrada y la clase diagnóstica. De acuerdo con Bishop (2006), los modelos de clasificación buscan aprender una función capaz de asignar una observación a una clase específica, procurando que dicha función no sólo se ajuste a los datos de entrenamiento, sino que mantenga capacidad de generalización frente a nuevos casos. En esta investigación, los seis algoritmos seleccionados permitieron comparar enfoques lineales, probabilísticos, basados en distancia, basados en márgenes, particiones del espacio y métodos de ensamble. **La regresión logística** es un modelo supervisado de clasificación probabilística utilizado para estimar la probabilidad de pertenencia de una observación a una clase determinada. En problemas binarios, como el abordado en esta investigación, el modelo estima la probabilidad de presencia o ausencia de cáncer de tiroides a partir de una combinación de las variables de entrada. Su salida puede interpretarse como una probabilidad condicional, lo cual resulta útil en contextos clínicos donde no sólo interesa asignar una clase, sino también observar el grado de confianza asociado a dicha asignación. Desde la perspectiva de Bishop (2006), este tipo de modelo pertenece a los enfoques discriminativos, ya que busca modelar directamente la relación entre las entradas y la clase, sin requerir una descripción completa de la distribución generativa de los datos.

Una de las principales ventajas de **la regresión logística** es su interpretabilidad y su comportamiento estable cuando la relación entre las variables y la clase puede aproximarse mediante una frontera de decisión relativamente simple. Sin embargo, esta misma característica representa una limitación cuando los patrones clínicos presentan relaciones no lineales o interacciones complejas entre variables. Por ello, en esta investigación se utilizó como una línea base inicial, útil para valorar si una aproximación lineal probabilística era suficiente para discriminar entre ambas clases diagnósticas. El **árbol de decisión** es un algoritmo supervisado que organiza el proceso de clasificación mediante una secuencia de reglas jerárquicas aplicadas sobre las variables de entrada. Cada división del árbol separa el espacio de características en regiones más específicas, hasta llegar a nodos terminales donde se asigna una clase. En este sentido, el modelo puede entenderse como una forma de particionar el espacio de datos en subconjuntos progresivamente más homogéneos respecto a la variable objetivo. Bajo el marco general de Bishop (2006), este tipo de aproximación se relaciona con el problema de construir una función de decisión capaz de separar observaciones pertenecientes a diferentes clases.

Su principal ventaja es que permite representar relaciones no lineales de manera intuitiva, mediante reglas que pueden ser interpretadas con relativa facilidad. Esto resulta relevante en datos clínicos, donde la posibilidad de observar condiciones de decisión explícitas puede aportar claridad al análisis. No obstante, **los árboles de decisión** pueden ser sensibles a pequeñas variaciones en los datos y tienden a sobreajustarse cuando crecen con demasiada profundidad. Por esta razón, su desempeño debe evaluarse considerando no sólo el ajuste al conjunto de entrenamiento, sino también su capacidad de generalización frente a datos no observados, aspecto central en la evaluación de modelos supervisados (Bishop, 2006). **Random Forest** es un método de ensamble que combina múltiples **árboles de decisión** con el propósito de obtener una clasificación más robusta que la producida por un único árbol. Para en lugar de depender de una sola estructura de decisión, el modelo integra varias particiones del espacio de características y genera una predicción final a partir del consenso entre los árboles individuales. Desde el enfoque de Bishop (2006), los métodos de ensamble se justifican porque la combinación de varios modelos puede mejorar la capacidad de generalización, especialmente cuando los modelos individuales presentan variabilidad o sensibilidad frente a los datos de entrenamiento.

Cumpliendo con lo establecido en esta tesis, **Random Forest** se incluyó como una línea base supervisada capaz de capturar relaciones no lineales y posibles interacciones entre variables clínicas. Su utilidad radica en que reduce parte de la inestabilidad asociada a los árboles individuales y permite construir una frontera de decisión más flexible. Sin embargo, aunque suele ofrecer un mejor rendimiento que un árbol aislado, también

puede disminuir la interpretabilidad global del modelo, debido a que la predicción depende del comportamiento conjunto de múltiples árboles. Por ello, su inclusión permitió comparar el desempeño de un enfoque de ensamble frente a modelos más simples y directamente interpretables.

SVM con kernel RBF es un algoritmo supervisado que busca construir una frontera de decisión con el mayor margen posible entre las clases. En su formulación con kernel, el modelo permite proyectar los datos a un espacio de mayor dimensionalidad sin realizar explícitamente dicha transformación, facilitando la separación de patrones que no son linealmente separables en el espacio original. Bishop (2006) describe los métodos basados en kernels como una estrategia para trabajar con fronteras de decisión no lineales mediante funciones que calculan similitudes entre observaciones. El kernel RBF resulta especialmente útil cuando la separación entre clases puede depender de relaciones locales y no lineales dentro del espacio de características. Para esta investigación, su uso permitió evaluar si los historiales clínicos presentaban una estructura de separación más compleja que la capturable por modelos lineales. Sin embargo, este tipo de modelo puede ser sensible a la selección de hiperparámetros y a la escala de las variables, por lo que requiere una preparación cuidadosa de los datos. Su incorporación como línea base permitió contrastar el rendimiento de un clasificador de margen máximo frente a modelos probabilísticos, basados en distancia y de ensamble.

KNN, configurado con k=5, es un algoritmo supervisado basado en la proximidad entre observaciones. A diferencia de los modelos que ajustan explícitamente una función paramétrica durante el entrenamiento, **KNN** clasifica una nueva instancia a partir de las clases de sus vecinos más cercanos en el espacio de características. En este caso, el valor k=5 indica que la predicción se determina considerando las cinco observaciones más próximas, lo que permite reducir la sensibilidad que tendría una decisión basada únicamente en el vecino más cercano. Bishop (2006) presenta los métodos de vecinos cercanos como una aproximación no paramétrica, donde la estructura de los datos se utiliza directamente para realizar la clasificación. Bajo el contexto de esta investigación, **KNN** permitió evaluar si los historiales clínicos de una misma clase tendían a ubicarse cerca entre sí dentro del espacio de características. Su principal ventaja es que no impone una forma específica sobre la frontera de decisión, lo cual puede ser útil cuando las relaciones entre variables son complejas. Sin embargo, su desempeño depende de la escala de las variables, de la métrica de distancia utilizada y de la distribución de los datos. Además, puede verse afectado cuando existen muchas dimensiones o cuando las clases se encuentran altamente solapadas. Por esta razón, su resultado fue interpretado como una medida de separabilidad local entre observaciones clínicas.

Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes y en el supuesto de independencia condicional entre las variables de entrada dada la clase. Bajo este enfoque, el modelo estima la probabilidad de que una observación pertenezca a cada clase y asigna la categoría con mayor probabilidad posterior. De acuerdo con Bishop (2006), los modelos probabilísticos permiten representar la incertidumbre asociada a la clasificación, lo cual resulta especialmente relevante en problemas donde las observaciones pueden presentar ruido, variabilidad o información incompleta. En esta investigación, **Naive Bayes** se utilizó como una línea base generativa para la clasificación binaria de los historiales clínicos. Su principal fortaleza es su simplicidad y eficiencia, ya que puede producir resultados competitivos incluso con conjuntos de datos limitados. Sin embargo, su supuesto de independencia entre variables puede resultar restrictivo en contextos clínicos, donde los atributos suelen estar relacionados entre sí. A pesar de esta limitación, su inclusión permitió comparar el desempeño de un modelo probabilístico simple frente a enfoques discriminativos, basados en distancia, márgenes y ensambles. En conjunto, estos seis algoritmos permitieron establecer una línea base supervisada amplia antes de implementar modelos de aprendizaje profundo. La comparación entre ellos resultó metodológicamente pertinente porque cada uno representa una forma distinta de aproximar el problema de clasificación binaria: **la regresión logística** desde una perspectiva probabilística discriminativa, **Naive Bayes** desde un enfoque probabilístico generativo, **KNN** desde la proximidad local, **SVM** desde la maximización del margen, el **árbol de decisión** desde la partición jerárquica del espacio y **Random Forest** desde la combinación de múltiples clasificadores. Así, la línea base permitió valorar si los patrones clínicos podían ser capturados mediante modelos tradicionales de aprendizaje automático y sirvió como punto de referencia para evaluar el aporte de arquitecturas más complejas.

2.2.3 Aprendizaje profundo y redes neuronales

El aprendizaje profundo (Deep Learning) emplea redes neuronales con múltiples capas ocultas para aprender representaciones jerárquicas de los datos (Goodfellow et al., 2016). Cada neurona recibe entradas, las pondera mediante parámetros aprendibles, aplica una función de activación no lineal y produce una salida. Esta composición sucesiva de transformaciones permite que las primeras capas capturen patrones simples, mientras que las capas posteriores integran dichas representaciones en estructuras de mayor complejidad. En consecuencia, las redes profundas resultan especialmente útiles cuando la relación entre las variables de entrada y la salida no puede representarse adecuadamente mediante modelos lineales o reglas explícitas. Dentro de este enfoque, las funciones de activación cumplen un papel fundamental, ya que introducen no linealidad en la red y permiten que el modelo aprenda relaciones complejas entre los datos. Sin estas funciones, la composición de múltiples capas equivaldría, en términos generales, a una transformación lineal, limitando la capacidad expresiva del modelo. Por ello, la selección de la función de activación influye directamente en la forma en que la red propaga información, ajusta sus parámetros y representa patrones durante el entrenamiento.

Las funciones de activación más relevantes para esta tesis son ReLU (Rectified Linear Unit) y GELU (Gaussian Error Linear Unit). ReLU se caracteriza por activar únicamente los valores positivos y anular los negativos, lo que favorece una propagación eficiente de la señal en redes profundas (Goodfellow et al., 2016). GELU, por su parte, pondera suavemente la entrada mediante la función de distribución acumulada gaussiana, lo que produce una activación menos abrupta que ReLU y permite conservar información de manera gradual en función del valor de entrada (Hendrycks & Gimpel, 2016). Esta característica ha favorecido su uso en arquitecturas modernas de lenguaje, particularmente en modelos derivados de BERT. Dentro de las arquitecturas densas, el Perceptrón Multicapa (MLP) es una red feed-forward donde cada neurona se conecta con todas las neuronas de la capa siguiente (Goodfellow et al., 2016). En este tipo de red, la información fluye desde la capa de entrada hacia la capa de salida a través de una o varias capas ocultas, sin conexiones recurrentes ni retroalimentación interna. Esta estructura permite transformar progresivamente las representaciones de entrada hasta obtener una salida asociada a una tarea específica, como la clasificación binaria. En las experimentaciones complementarias de esta tesis, se diseñó un MLP cuya arquitectura replica parcialmente la estructura interna del bloque feed-forward utilizado en modelos tipo Transformer. Específicamente, se empleó una expansión dimensional de 768 a 3072 unidades, seguida de una activación GELU, una compresión de 3072 a 768 unidades, regularización mediante Dropout al 10% y una salida binaria. Esta configuración permitió que el clasificador operara sobre un espacio matemático compatible con las representaciones generadas por el modelo de lenguaje, manteniendo una relación estructural con la dimensión oculta empleada en arquitecturas tipo BERT (Devlin et al., 2019). En la figura 2.4 es posible observar la representación de una red neuronal emplead en aprendizaje profundo.

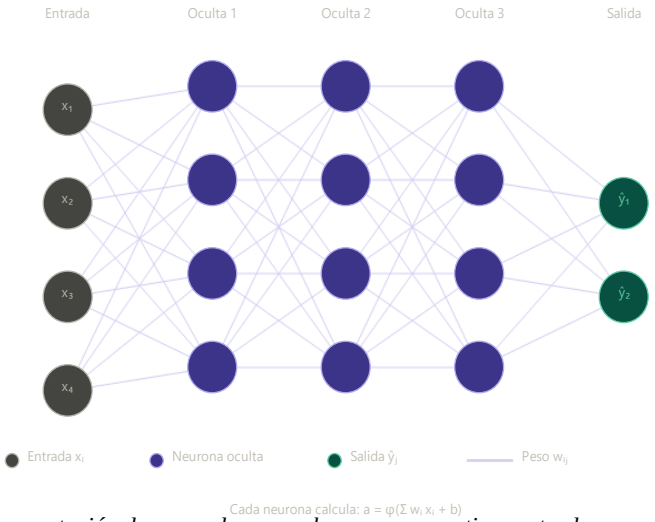

Figura 2.4: Representación de una red neuronal con sus respectivas entradas, capas ocultas y salida

2.3 Arquitectura Transformer y modelos de lenguaje preentrenados

2.3.1 Tokenización

La *tokenización* segmenta el texto en unidades discretas (*tokens*) que el modelo procesa. Los modelos modernos emplean algoritmos de subpalabras como *WordPiece* o *BPE* (*Byte-Pair Encoding*) que descomponen las palabras en fragmentos aprendidos estadísticamente del corpus de entrenamiento (Jurafsky & Martin, 2026). La elección del *tokenizador* tiene implicaciones directas sobre el rendimiento, como ilustra la Figura 2.5.

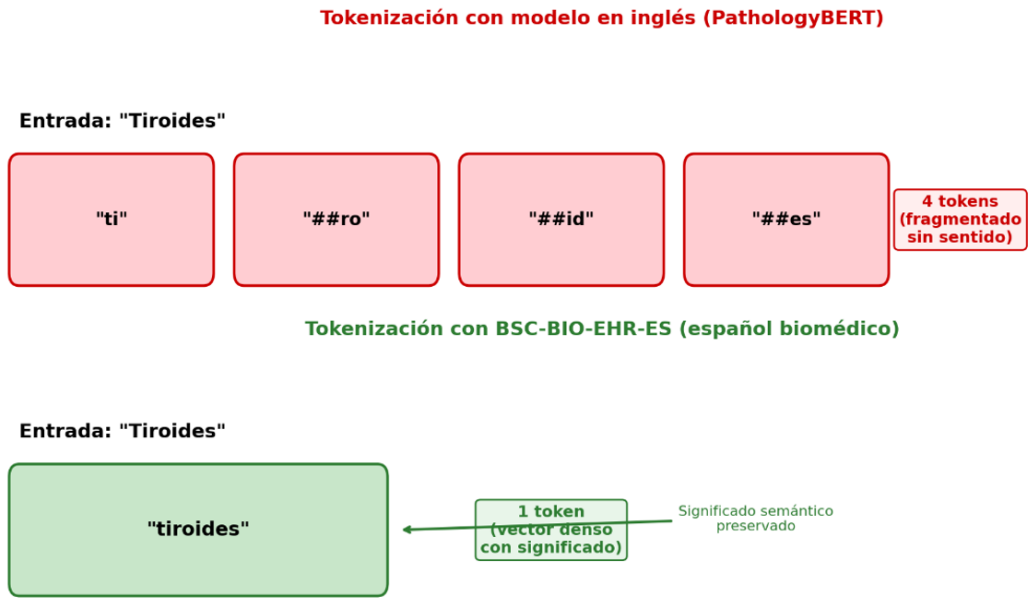

Figura 2.5: Impacto de la tokenización. Un tokenizador anglófono fragmenta "Tiroides" en 4 tokens sin sentido, mientras que BSC-BIO-EHR-ES lo representa como un único token denso.

Como se observa en la Figura 2.3, un *tokenizador* optimizado para inglés fragmenta de forma ineficiente el vocabulario médico en español, obligando al mecanismo de atención a invertir recursos en reconstruir el significado de palabras que un *tokenizador* nativo representaría como un único *token* (Jurafsky & Martin, 2026). Esta ineficiencia constituyó uno de los argumentos centrales para descartar modelos biomédicos anglófonos y favorecer *BSC-BIO-EHR-ES* con su vocabulario de 52,000 *tokens* médicos en español (Carrino et al., 2022).

2.3.2 La arquitectura Transformer

La arquitectura *Transformer*, propuesta por *Vaswani* et al. (Vaswani et al., 2017) en 2017 en el artículo seminal "*Attention Is All You Need*", representó un punto de inflexión en el PLN. A diferencia de las arquitecturas secuenciales precedentes, procesa la secuencia completa de entrada simultáneamente mediante autoatención (*Self-Attention*), capturando dependencias de largo alcance y facilitando la paralelización (Vaswani et al., 2017). La Figura 2.6 presenta la arquitectura del bloque codificador.

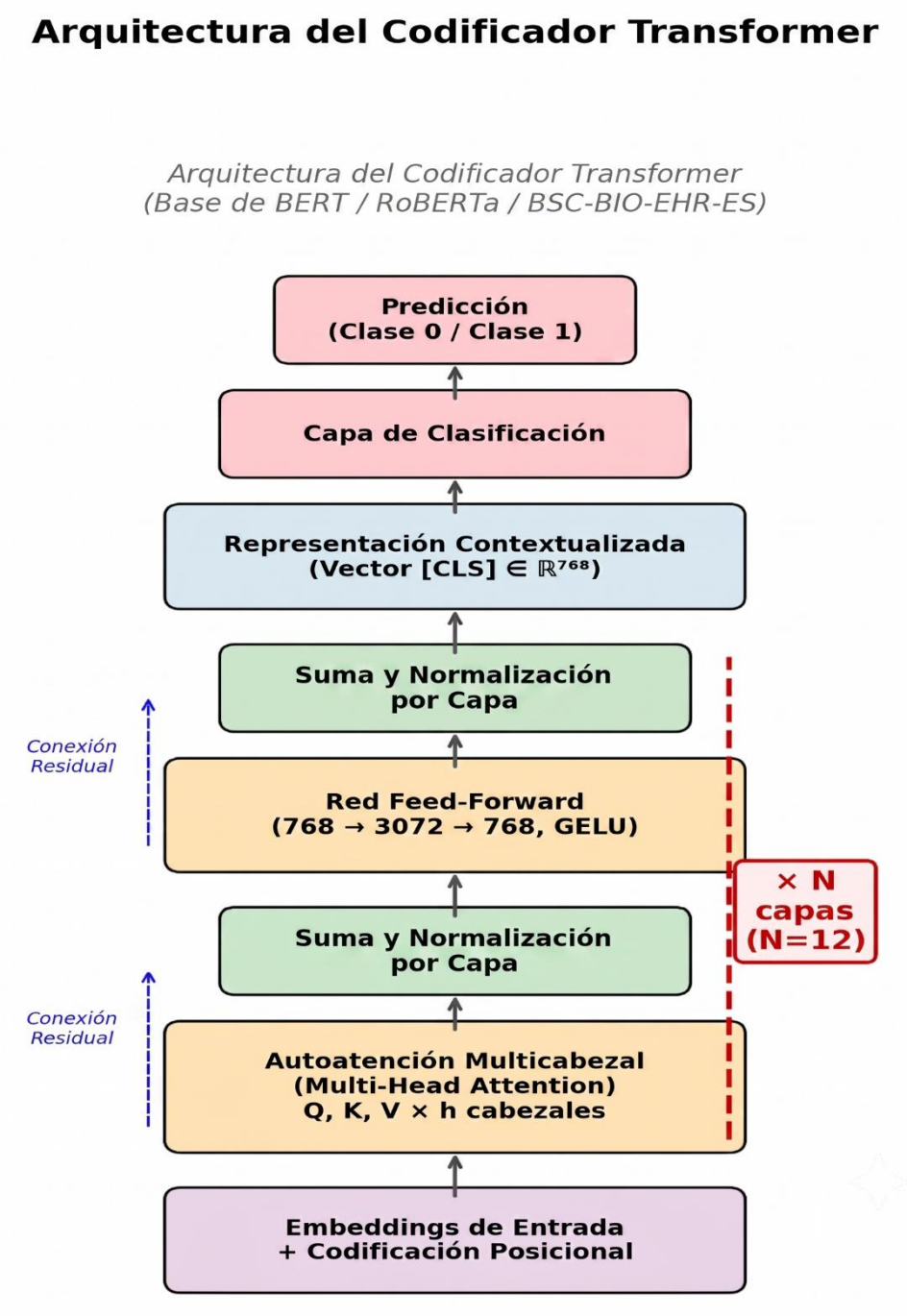

Figura 2.6: Arquitectura del bloque codificador del Transformer. Cada bloque consta de autoatención multicabezal y una red feed-forward, con conexiones residuales y normalización de capa. BSC-BIO-EHR-ES apila N=12 bloques.

2.3.2.1 Mecanismo de autoatención

Para cada token, el mecanismo calcula tres vectores: *Query (Q)*, *Key (K)* y *Value (V)*, obtenidos mediante matrices de pesos aprendibles. La atención se formaliza en la Ecuación 2.1.

$$Attention(Q, K, V) = softmax(\frac{QK^{T}}{\sqrt{d_{k}}}) \cdot V \quad (2.1)$$

donde **Q** es la matriz de consultas, **K** la de claves, **V** la de valores y $\mathbf{d_k}$ la dimensión de las claves, que actúa como factor de escala para evitar gradientes desvaneciente (Vaswani et al., 2017). El *Transformer* emplea h=12 cabezales de atención en paralelo, permitiendo atender simultáneamente a diferentes tipos de relaciones lingüísticas en distintos subespacios de representación.

2.3.3 Representación textual (estáticos → contextuales)

Para que un algoritmo computacional pueda procesar texto, es necesario transformar las secuencias lingüísticas en representaciones numéricas. Los *embeddings* o representaciones vectoriales densas codifican cada unidad lingüística como un vector de dimensionalidad reducida y fija en un espacio continuo, donde la proximidad geométrica se correlaciona con la similitud semántica (Jurafsky & Martin, 2026). A diferencia de los enfoques dispersos clásicos que generan vectores de alta dimensionalidad con la mayoría de valores en cero, los modelos contextualizados modernos producen representaciones que dependen dinámicamente del contexto, generando vectores diferentes para la misma palabra según la oración en que aparece (Devlin et al., 2019). La Figura 2.7 ilustra esta diferencia estructural.

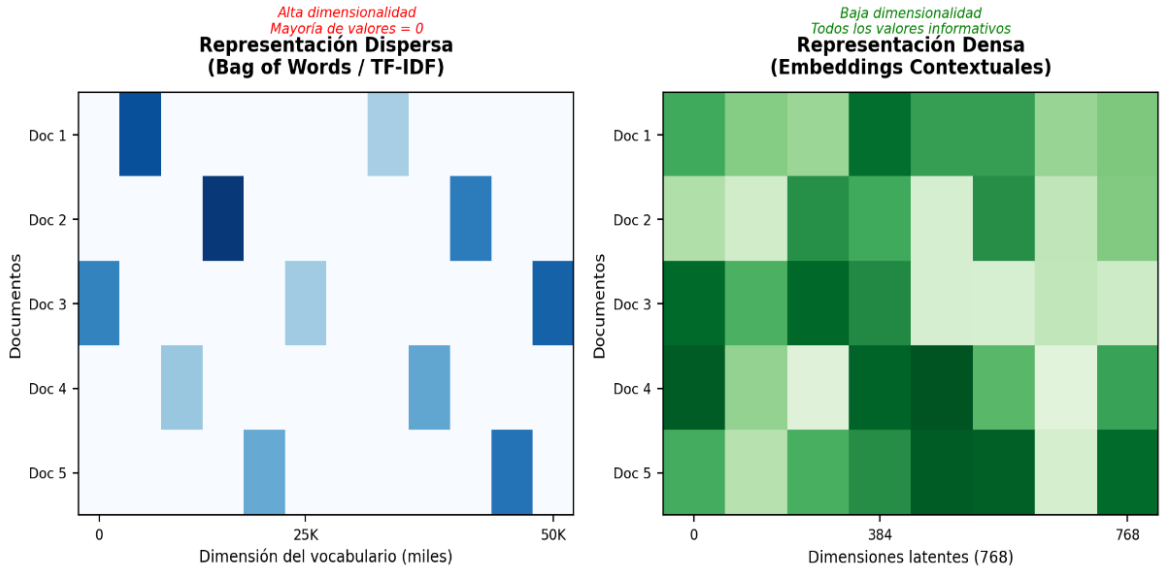

Figura 2.7: Comparación entre representaciones textuales dispersas (izquierda), donde la mayoría de valores son cero, y representaciones densas tipo embedding (derecha), donde todas las dimensiones codifican información semántica útil.

Mientras las representaciones dispersas requieren vectores de miles de dimensiones con información concentrada en pocas posiciones, los *embeddings* densos condensan el significado semántico en un espacio de dimensionalidad fija. En esta investigación, el modelo *BSC-BIO-EHR-ES* genera vectores densos de dimensión $d = 768$ a partir del token especial *[CLS]* (Carrino et al., 2022), que encapsula la representación semántica global de cada historia clínica procesada. Esta representación constituye la base sobre la que operan tanto el clasificador directo como el sistema complementario de aumento de datos en espacio latente.

2.3.4 Modelos de lenguaje especializados

2.3.4.1 BERT

BERT (Bidirectional Encoder Representations from Transformers), propuesto por Devlin et al. (2019), es un modelo de lenguaje preentrenado basado en la arquitectura Transformer, diseñado para aprender representaciones bidireccionales profundas a partir de texto no etiquetado. A diferencia de los modelos unidireccionales, que procesan la información siguiendo únicamente el contexto izquierdo o derecho, BERT permite que cada token sea representado considerando simultáneamente la información anterior y posterior dentro de la secuencia. Esta característica resulta especialmente relevante en tareas de comprensión del lenguaje, donde el significado de una palabra o fragmento textual depende del contexto completo en el que aparece. El preentrenamiento de BERT se realiza mediante dos objetivos principales: modelo de lenguaje enmascarado (Masked Language Modeling, MLM) y predicción de la siguiente oración (Next Sentence Prediction, NSP). En el MLM, se selecciona aleatoriamente un porcentaje de tokens de entrada, equivalente al 15%, y el modelo debe predecir los tokens originales utilizando el contexto disponible. Este procedimiento permite que BERT aprenda relaciones semánticas y sintácticas de manera bidireccional, ya que la predicción no depende únicamente de los tokens anteriores, sino también de los posteriores. En consecuencia, el modelo desarrolla representaciones contextuales más ricas que las obtenidas mediante enfoques secuenciales tradicionales.

Por su parte, la tarea NSP consiste en determinar si dos segmentos de texto aparecen de forma consecutiva dentro del corpus original. Este objetivo fue incorporado para favorecer el aprendizaje de relaciones entre oraciones, lo cual resulta útil en tareas donde no sólo importa el significado de tokens individuales, sino también la relación discursiva entre fragmentos textuales. Aunque en modelos posteriores esta tarea ha sido revisada o reemplazada, en la formulación original de BERT constituye uno de los componentes centrales del proceso de preentrenamiento descrito por Devlin et al. (2019).

2.3.4.2 RoBERTa

RoBERTa (Robustly Optimized BERT Approach), propuesto por Liu et al. (2019), mantiene la arquitectura general de BERT, pero modifica y optimiza su procedimiento de preentrenamiento. Su planteamiento parte de una revisión crítica de BERT, señalando que parte de su desempeño dependía no sólo de la arquitectura Transformer, sino también de decisiones relacionadas con la cantidad de datos, el tamaño de los lotes, la duración del entrenamiento, el uso de secuencias más largas y la estrategia de enmascaramiento. En este sentido, RoBERTa no propone una arquitectura completamente distinta, sino una versión robustamente optimizada del proceso de entrenamiento. Una de las principales modificaciones de RoBERTa consiste en eliminar la tarea de predicción de la siguiente oración (Next Sentence Prediction, NSP). A diferencia de BERT, que combina el modelo de lenguaje enmascarado con NSP durante el preentrenamiento, RoBERTa conserva únicamente el objetivo de modelado de lenguaje enmascarado, al observar que la eliminación de NSP podía mejorar el desempeño en diversas tareas de comprensión del lenguaje. Esta modificación permitió concentrar el entrenamiento en la recuperación contextual de tokens enmascarados, fortaleciendo la capacidad del modelo para construir representaciones bidireccionales a partir del contenido textual.

Otra diferencia relevante es el uso de enmascaramiento dinámico. Mientras que en BERT los patrones de enmascaramiento podían mantenerse de forma relativamente fija durante el entrenamiento, RoBERTa genera diferentes patrones de máscara a lo largo del proceso. Esto permite que el modelo observe variaciones más amplias del mismo corpus y aprenda representaciones menos dependientes de una configuración específica de tokens ocultos. En consecuencia, el entrenamiento se vuelve más diverso y puede favorecer una mejor generalización en tareas posteriores.

2.3.4.3 *BSC-BIO-EHR-ES*

El modelo BSC-BIO-EHR-ES, desarrollado por el Barcelona Supercomputing Center dentro del Plan de Tecnologías del Lenguaje (Carrino et al., 2022), constituye una especialización de RoBERTa para el dominio

biomédico-clínico en español. Este modelo fue diseñado para responder a una limitación relevante dentro del procesamiento de lenguaje natural biomédico: la escasez de modelos especializados en español capaces de representar adecuadamente textos clínicos, notas médicas y documentos derivados de historiales electrónicos de salud. Aunque los modelos generales de lenguaje pueden capturar estructuras gramaticales y semánticas amplias, su desempeño puede verse limitado cuando se aplican a dominios altamente especializados, donde el vocabulario, las abreviaturas, las relaciones terminológicas y las formas narrativas difieren del lenguaje común.

BSC-BIO-EHR-ES parte de la arquitectura RoBERTa y adapta su entrenamiento al contexto biomédico-clínico mediante el uso de corpus especializados en español. Esta adaptación resulta metodológicamente pertinente para la presente investigación, debido a que los historiales clínicos contienen información médica expresada en lenguaje natural, con estructuras variables, términos técnicos, menciones anatómicas, antecedentes, síntomas, hallazgos diagnósticos y descripciones clínicas que requieren una representación contextual más precisa que la ofrecida por enfoques tradicionales basados únicamente en frecuencia o coincidencia léxica. A diferencia de los modelos entrenados sobre textos generales, BSC-BIO-EHR-ES fue expuesto durante su preentrenamiento a documentos biomédicos y clínicos, lo que le permite construir representaciones más cercanas al lenguaje utilizado en escenarios médicos. Esta característica es especialmente importante en tareas de clasificación diagnóstica, ya que el significado de una expresión clínica no depende únicamente de palabras aisladas, sino de su relación con otros elementos del historial. Por ejemplo, la presencia de un término asociado a una enfermedad puede tener implicaciones distintas si aparece como antecedente, sospecha diagnóstica, negación, hallazgo confirmado o referencia familiar. La Tabla 2.2 presenta sus especificaciones técnicas.

Tabla 2.2: Especificaciones técnicas del modelo BSC-BIO-EHR-ES

Parámetro	**Especificación**
Identificador	*PlanTL-GOB-ES/bsc-bio-ehr-es*
Arquitectura base	*RoBERTa* (solo codificador)
Dominio de preentrenamiento	Biomédico + *EHR* en español
Capas / *Hidden Size* / *Heads*	12 / 768 / 12
Intermediate Size	3,072 (expansión 4×)
Función de activación	*GELU*
Vocabulario	~52,000 tokens médicos en español
Parámetros totales	~125 millones
Corpus de preentrenamiento	>95 millones de tokens de *EHR* reales

El factor diferencial de *BSC-BIO-EHR-ES* reside en su preentrenamiento con más de 95 millones de *tokens* de Registros Electrónicos de Salud reales en español (Carrino et al., 2022), lo que le confiere capacidad de discernimiento contextual clínico. La Tabla 2.3 muestra su superioridad en *benchmarks* de dominio clínico frente a modelos alternativos.

Tabla 2.3: Comparativa de rendimiento (F1-Score) en tareas de dominio clínico

Modelo	**Arquitectura**	**CANTEMIST**	**ICTUSnet**
bsc-bio-ehr-es	*RoBERTa* (*Bio+EHR*)	0.834	0.8756
XLM-R-Galén	*XLM-R* (Multilingüe)	0.8078	0.8716
BETO-Galén	*BERT* (General Español)	0.8153	0.8498
mBERT	*BERT* (Multilingüe)	0.8116	0.8631
roberta-base-bne	*RoBERTa* (General)	0.7875	0.8677

La Tabla 2.3 evidencia la superioridad consistente de *BSC-BIO-EHR-ES*, alcanzando el *F1-Score* más alto tanto en detección oncológica (*CANTEMIST*: 0.8340) como en textos clínicos reales (*ICTUSnet*: 0.8756), lo que sustentó su selección como modelo base para el ajuste fino (Carrino et al., 2022).

2.3.5 Aprendizaje por transferencia y ajuste fino

El aprendizaje por transferencia (*Transfer Learning*) reutiliza un modelo preentrenado en una tarea general como punto de partida para una tarea específica (Devlin et al., 2019). En el *PLN* moderno, el esquema predominante consiste en un preentrenamiento no supervisado sobre corpus masivos seguido de una adaptación supervisada (Devlin et al., 2019). La Figura 2.8 sintetiza este paradigma aplicado al flujo principal de esta tesis.

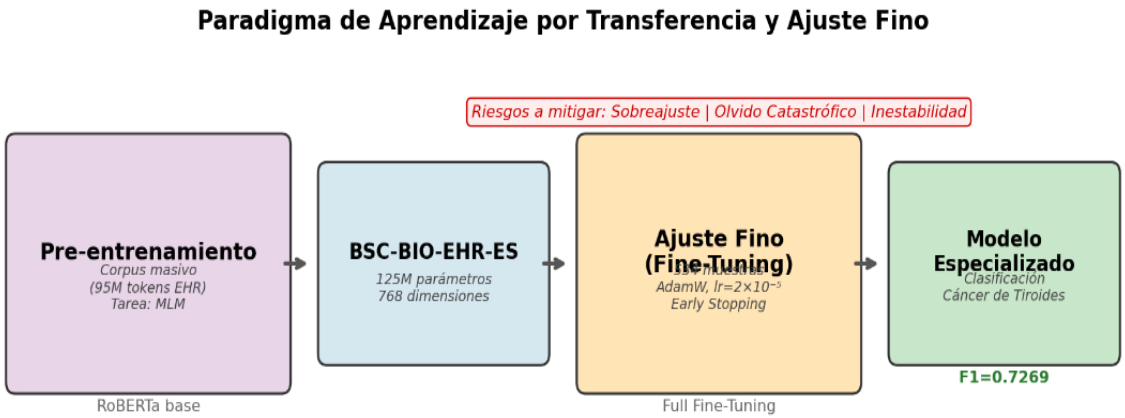

Figura 2.8: Paradigma de aprendizaje por transferencia. BSC-BIO-EHR-ES, preentrenado con 95M tokens de EHR, se somete a ajuste fino completo con 534 muestras bajo configuración regularizada.

La Figura 2.4 muestra los riesgos del ajuste fino con *Datasets* reducidos: el sobreajuste (*overfitting*) y el olvido catastrófico (*catastrophic forgetting*) (Goodfellow et al., 2016), donde el entrenamiento intensivo degrada el conocimiento previamente adquirido. Estos riesgos se mitigan con las técnicas de regularización descritas a continuación.

2.3.6 Técnicas de regularización

La regularización comprende estrategias para prevenir el sobreajuste, favoreciendo la generalización (Goodfellow et al., 2016). En esta tesis se emplearon tres técnicas complementarias para estabilizar el aprendizaje en un régimen de datos escasos (*N=534*):

Early Stopping detiene el entrenamiento cuando la métrica de validación deja de mejorar durante un número predeterminado de épocas (paciencia = 2), impidiendo la memorización más allá del punto óptimo (Goodfellow et al., 2016).

Weight Decay es una regularización L2 desacoplada implementada en el optimizador *AdamW* (valor = 0.01), que penaliza pesos de gran magnitud añadiendo un término proporcional a su norma cuadrada (Goodfellow et al., 2016).

Dropout desactiva aleatoriamente el 10% de las neuronas durante cada paso de entrenamiento, forzando representaciones redundantes y robustas que no dependan de neuronas individuales (Goodfellow et al., 2016).

2.4 Reducción de dimensionalidad

Los datos generados por *BSC-BIO-EHR-ES* habitan en un espacio vectorial de 768 dimensiones, lo que imposibilita su inspección visual directa. Para auditar la calidad de los datos sintéticos y verificar la coherencia topológica en las experimentaciones complementarias, se emplean dos técnicas de reducción de dimensionalidad, cuyas diferencias se ilustran en la Figura 2.9.

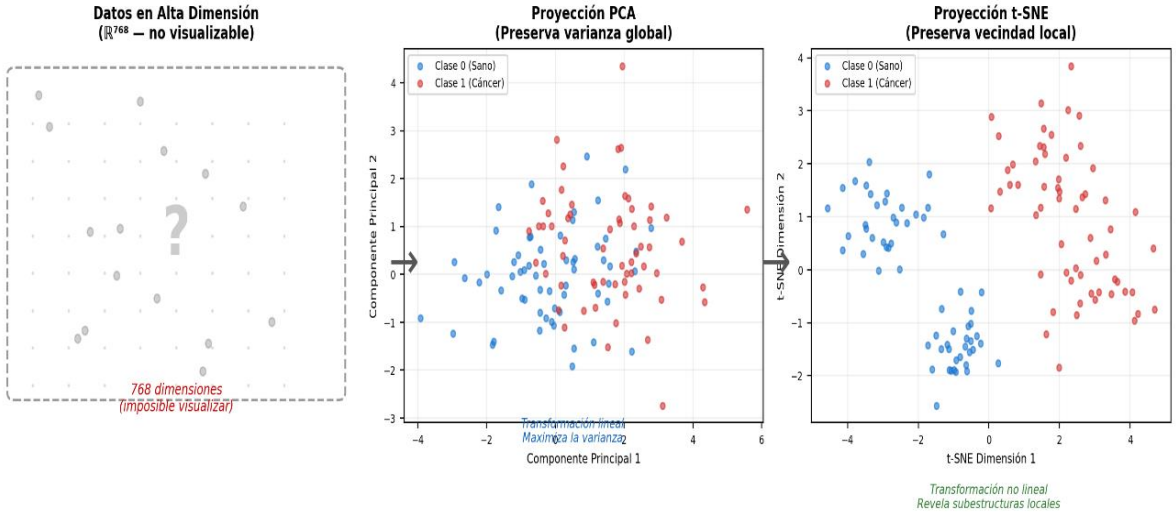

Figura 2.9: Comparación conceptual entre PCA y t-SNE. Desde $\mathbb{R}^{768}$ (izquierda), PCA preserva varianza global (centro) mientras que t-SNE revela subestructuras locales (derecha).

2.4.1 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (*PCA*) es una técnica de reducción de dimensionalidad lineal que transforma variables correlacionadas en un nuevo sistema de coordenadas ortogonales denominadas componentes principales, ordenadas por la varianza que capturan (Bishop, 2006). Matemáticamente, dado un conjunto de datos centrado $X \in R^{n \times d}$, *PCA* proyecta estos datos resolviendo el problema de valores propios de la matriz de covarianza Σ:

$$\Sigma v = \lambda v \qquad (2.2)$$

Donde $\Sigma = \frac{1}{n-1} X^T X$,el vector v representa los vectores propios (las direcciones de las componentes principales) y λ corresponde a los valores propios, los cuales indican la magnitud de la varianza explicada por cada componente.

Esta sección se utiliza como inicialización para *t-SNE* (proporcionando un punto de partida determinista y más estable frente a la convergencia) y como herramienta de visualización exploratoria del *Dataset* previo a los algoritmos de agrupamiento, ayudando a preservar la estructura geométrica global de las representaciones clínicas.

2.4.2 t-SNE

Propuesta por *Van der Maaten* y *Hinton* (Maaten & Hinton, 2008), es una técnica de reducción no lineal diseñada específicamente para la visualización de datos de alta dimensionalidad. A diferencia de *PCA*, que preserva la varianza global, *t-SNE* se optimiza para preservar las relaciones de vecindad local: los puntos similares en el espacio original permanecen cercanos en la proyección bidimensional (Maaten & Hinton, 2008).

El algoritmo opera en dos fases fundamentales. Primero, construye una distribución de probabilidad sobre pares de puntos en el espacio original, asignando una alta probabilidad a pares similares. Para dos puntos x_i y x_j, la similitud condicional se define mediante una distribución *gaussiana*:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \quad (2.3)$$

Para garantizar la simetría y eficiencia computacional, la probabilidad conjunta en el espacio de alta dimensión se define como:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \quad (2.4)$$

donde la varianza σ_i^2 se ajusta para que la distribución alcance un nivel de perplejidad (*perplexity*) definido por el usuario. Posteriormente, *t-SNE* define una distribución análoga Q en el espacio reducido para los puntos proyectados y_i e y_j. Para esto, utiliza una distribución t de *Student* con un grado de libertad (equivalente a una distribución de *Cauchy*):

$$q_{ij} = \frac{\left(1+||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l}(1+||y_k - y_l||^2)^{-1}} \quad (2.5)$$

Finalmente, el algoritmo optimiza las ubicaciones en el espacio de baja dimensión minimizando la divergencia de *Kullback-Leibler* (*KL*) entre las distribuciones *P* y *Q* mediante descenso de gradiente:

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (2.6)$$

La pesada cola de la distribución *t* de *Student* en el espacio de baja dimensión mitiga el problema del apiñamiento (*crowding problem*), permitiendo que los puntos disímiles se alejen más y logrando una separación visual mucho más clara entre clústeres.

La Figura 2.9 ilustra cómo *PCA*, al ser lineal, produce proyecciones donde los clústeres pueden solaparse, mientras que *t-SNE* revela subgrupos y zonas de solapamiento críticas para la validación de datos sintéticos.

2.5 Aumento de datos en espacio latente

Como parte de las experimentaciones complementarias orientadas a mitigar el cuello de botella por escasez de datos (*N=534*), se implementó un enfoque de aumento de datos en espacio latente que opera directamente sobre los *embeddings* de 768 dimensiones (Carrino et al., 2022). La Figura 2.10 presenta el flujo completo de esta arquitectura híbrida.

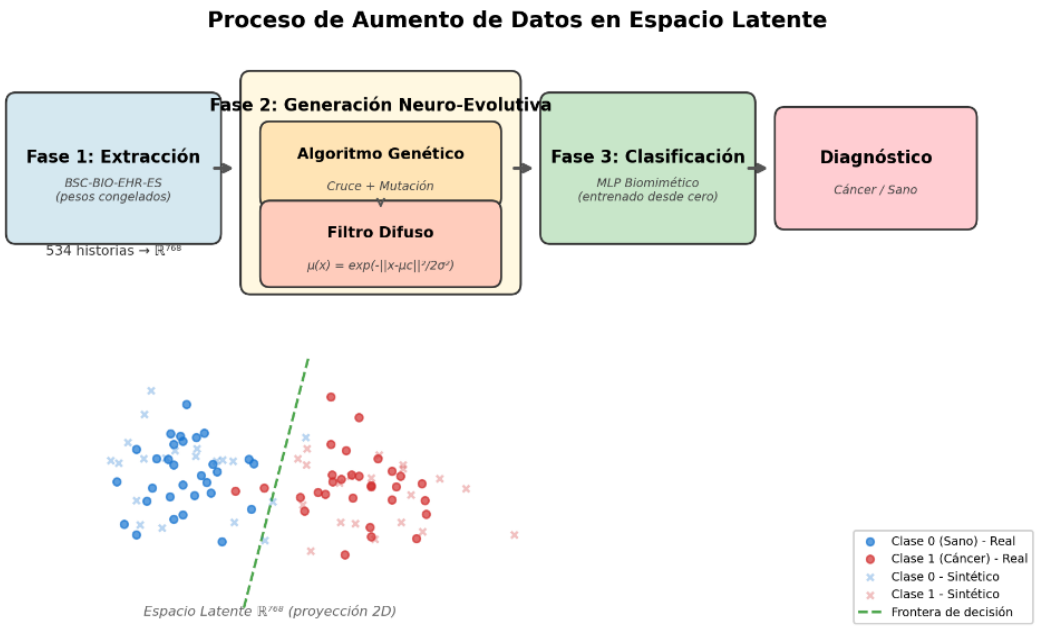

Figura 2.10: Arquitectura del sistema de aumento de datos en espacio latente. Las tres fases (extracción congelada, generación neuro-evolutiva y clasificación MLP) se presentan arriba. Abajo, proyección 2D con datos reales y sintéticos.

2.5.1 Algoritmos genéticos

Los algoritmos genéticos son metaheurísticas de optimización inspiradas en la evolución biológica, formalizados por Holland (1975). Su funcionamiento se basa en la representación de posibles soluciones como individuos dentro de una población, los cuales evolucionan mediante operadores de selección, cruce y mutación. A través de este proceso, las soluciones con mayor aptitud tienen mayor probabilidad de conservarse y transmitir sus características a nuevas generaciones, favoreciendo progresivamente la exploración de regiones prometedoras del espacio de búsqueda. En esta tesis, cada individuo fue representado como un vector sintético, compatible con la dimensionalidad de las representaciones generadas por el modelo de lenguaje. Esta decisión permitió que los individuos producidos por el algoritmo genético operaran dentro del mismo espacio vectorial utilizado por el clasificador, evitando una ruptura entre la representación semántica original y los datos sintéticos generados. De esta manera, cada vector sintético funcionó como una posible instancia artificial dentro del espacio latente aprendido por el modelo.

El operador de cruce empleado fue cxBlend con (alpha = 0.5), aplicado entre dos vectores padres pertenecientes a la misma clase. Este operador genera nuevos individuos mediante una interpolación lineal entre ambos vectores, permitiendo producir representaciones intermedias que conservan información de los dos padres. La restricción de cruzar individuos de la misma clase tuvo como finalidad preservar la coherencia semántica y diagnóstica de los vectores sintéticos, evitando combinaciones entre representaciones asociadas a clases distintas que pudieran introducir ambigüedad en el conjunto aumentado.

2.5.2 Lógica difusa como control de calidad

La lógica difusa, propuesta por *Zadeh* (Zadeh, 1965), permite trabajar con grados de verdad parciales en el intervalo [0, 1]. En esta investigación, se emplea como mecanismo de control de calidad de los vectores sintéticos. El grado de pertenencia de un vector candidato x a la clase c se define mediante la Ecuación 2.7.

$\mu c(x) = exp(\frac{-||x - \mu c||^2}{2\sigma c^2})$ (2.7)

Donde μc es el centroide de los vectores reales y σc la dispersión media (Zadeh, 1965). Esta función penaliza exponencialmente a los *outliers*, garantizando que solo sobrevivan vectores estadísticamente coherentes con los pacientes reales.

2.6 Evaluación y validación experimental

2.6.1 Validación cruzada

La validación cruzada estratificada de K pliegues divide el conjunto de datos en K subconjuntos mutuamente excluyentes, entrenando el modelo K veces con K-1 pliegues y validando con el pliegue restante. La estratificación permite preservar la proporción original de clases en cada partición, lo cual resulta especialmente importante en problemas de clasificación binaria donde puede existir desbalance entre las categorías analizadas. Desde la perspectiva de Bishop (2006), este tipo de procedimientos permite estimar con mayor estabilidad la capacidad de generalización de un modelo, al evaluar su desempeño sobre datos no utilizados durante el entrenamiento.

Dentro de la investigación, la validación cruzada fue empleada para reducir la dependencia de una única partición entrenamiento-prueba. Dado que el corpus disponible estuvo conformado por N=534 historiales clínicos, una división simple de los datos podía producir resultados sensibles a la selección aleatoria de casos. Por ello, el uso de K pliegues permitió aprovechar de manera más eficiente el conjunto disponible, ya que cada observación fue utilizada tanto para entrenamiento como para validación en diferentes iteraciones, sin que un mismo caso formara parte simultáneamente del entrenamiento y la evaluación dentro de un mismo pliegue.

La variante anidada, conocida como Nested Cross-Validation, incorpora un segundo bucle interno dentro de cada partición externa. El bucle externo se utiliza para estimar el desempeño general del modelo, mientras que el bucle interno se reserva para decisiones relacionadas con el ajuste del entrenamiento, como la selección de hiperparámetros o el criterio de early stopping. Esta separación es metodológicamente relevante porque evita que los datos utilizados para evaluar el desempeño final influyan en el proceso de optimización del modelo, reduciendo así el sesgo optimista asociado al uso repetido de los mismos datos para ajustar y evaluar (Bishop, 2006).

2.6.2 Métricas de evaluación

La evaluación del rendimiento de un clasificador binario en contexto clínico requiere métricas que capturen distintas dimensiones del comportamiento. Los costos de los falsos positivos y falsos negativos difieren sustancialmente, por lo que la selección de métricas tiene implicaciones bioéticas directas (Bishop, 2006). Las métricas centrales se formalizan a continuación.

$Precisión\ (P) = \frac{VP}{(VP + FP)}$ *(2.8)*

Donde *VP* son verdaderos positivos y *FP* falsos positivos. Una alta precisión minimiza alarmas falsas, protegiendo la integridad psicológica del paciente (principio de no maleficencia).

$Sensibilidad\ (R) = \frac{VP}{(VP + FN)}$ (2.9)

Donde *FN* son falsos negativos. Una alta sensibilidad garantiza que ningún caso real pase desapercibido (principio de beneficencia).

$$F1\text{-}Score = 2 \cdot \frac{(P \cdot R)}{(P + R)} \qquad (2.10)$$

El *F1-Score*, media armónica entre precisión y sensibilidad, fue establecido como métrica principal de éxito (Bishop, 2006).

$$MCC = \frac{(P \cdot R)\,(VP \cdot VN - FP \cdot FN)}{\sqrt{((VP+FP)(VP+FN)(VN+FP)(VN+FN))}} \qquad (2.11)$$

El Coeficiente de Correlación de *Matthews* (*MCC*) considera las cuatro celdas de la matriz de confusión, proporcionando una medida equilibrada especialmente informativa en *Datasets* balanceados (Bishop, 2006). La Tabla 2.4 sintetiza todas las métricas empleadas, incluyendo las de calibración.

Tabla 2.4: Métricas de evaluación, definición y relevancia clínica.

Métrica	Definición	Relevancia Clínica	Principio / Propósito
Precisión	$\frac{\text{VP}}{(\text{VP} + \text{FP})}$	Minimiza alarmas falsas	No maleficencia
Sensibilidad	$\frac{\text{VP}}{(\text{VP} + \text{FN})}$	Evita omisión de diagnósticos	Beneficencia
F1-Score	$2 \cdot \frac{(\text{P} \cdot \text{R})}{(\text{P} + \text{R})}$	Balance detección/confianza	Equilibrio bioético
MCC	Correlación de *Matthews*	Evaluación equilibrada	Robustez estadística
ROC-AUC	Área bajo curva *ROC*	Capacidad discriminativa	Independencia de umbral
PR-AUC	Área bajo curva *P-R*	Calidad sostenida	Confiabilidad clínica
Brier Score	Media de $(\hat{p}_i - y_i)^2$	Exactitud probabilística	Calibración global
ECE	Error de Calibración Esperado	Fidedigna de confianzas	Integridad probabilística

La Tabla 2.4 incluye el *Brier Score*, que mide la exactitud de las probabilidades predichas como el error cuadrático medio entre la probabilidad estimada y el resultado real; y el *ECE* (*Expected Calibration Error*), que cuantifica la discrepancia entre la confianza del modelo y su precisión real (Bishop, 2006). Un *ECE* bajo indica que cuando el modelo predice un 80% de probabilidad, efectivamente acierta el 80% de las veces.

2.6.3 Pruebas estadísticas no paramétricas

La validación estadística de los resultados experimentales constituye un requisito indispensable para determinar si las diferencias observadas entre métodos de clasificación son genuinas o producto del azar (Demšar, 2006). En esta investigación, la comparación del modelo BSC-BIO-EHR-ES frente a los seis algoritmos clásicos se realizó mediante dos pruebas estadísticas no paramétricas.

El uso de pruebas no paramétricas respondió a la naturaleza de los resultados obtenidos durante la validación cruzada, ya que no se asumió normalidad en la distribución de las métricas ni igualdad de varianzas entre los modelos comparados. En escenarios de aprendizaje automático, especialmente cuando se trabaja con conjuntos de datos limitados y múltiples particiones de evaluación, las diferencias de desempeño pueden variar entre pliegues y no necesariamente seguir una distribución normal. Por esta razón, las pruebas no paramétricas ofrecen una alternativa más robusta para contrastar el rendimiento relativo entre clasificadores.

2.6.3.1 Test de Friedman

El test de Friedman es una prueba no paramétrica para diseños de medidas repetidas que evalúa si existen diferencias estadísticamente significativas entre tres o más tratamientos relacionados. No requiere el supuesto de normalidad, lo que lo hace apropiado para comparar métricas de rendimiento de múltiples clasificadores sobre las mismas particiones de datos (Demšar, 2006). El procedimiento asigna rangos a los tratamientos dentro de cada fold y calcula un estadístico chi-cuadrado bajo la hipótesis nula de rendimiento idéntico. Un p-value inferior a (alpha = 0.05) indica que al menos un clasificador difiere significativamente, justificando pruebas post-hoc (Demšar, 2006).

El test de Friedman se empleó como una prueba global para comparar el desempeño del modelo BSC-BIO-EHR-ES frente a los algoritmos clásicos utilizados como línea base. Para cada partición de evaluación, los clasificadores fueron ordenados de acuerdo con la métrica analizada, asignando mejores rangos a los modelos con mayor rendimiento. Posteriormente, se calcularon los rangos promedio de cada clasificador con el fin de determinar si las diferencias observadas podían atribuirse únicamente a variaciones aleatorias o si existía evidencia estadística de un comportamiento diferencial entre los métodos. La hipótesis nula del test estableció que todos los clasificadores presentaban un rendimiento equivalente. Bajo esta hipótesis, las diferencias entre los rangos promedio deberían ser pequeñas y compatibles con el azar. Por el contrario, la hipótesis alternativa sostuvo que al menos uno de los modelos evaluados presentaba un desempeño significativamente distinto. En este sentido, el test de Friedman no identifica por sí mismo qué pares de clasificadores difieren entre sí, sino que funciona como una prueba inicial para determinar si existe una diferencia global dentro del conjunto de modelos comparados.

$$\chi_F^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1) \quad (2.12)$$

2.6.3.2 Test de rangos con signo de Wilcoxon

El test de Wilcoxon es una prueba no paramétrica para comparar dos muestras relacionadas, utilizada como prueba post-hoc para comparaciones pareadas entre el modelo propuesto y cada algoritmo de la línea base (Demšar, 2006). Opera sobre las diferencias pareadas de las métricas en cada fold: las ordena por valor absoluto, asigna rangos y compara la suma de rangos positivos contra la de negativos (Demšar, 2006). Un p-value significativo indica que existe evidencia estadística para rechazar la hipótesis nula de igualdad de rendimiento entre ambos modelos. En esta investigación, se aplicó sobre F1-Score, ROC-AUC y PR-AUC utilizando las observaciones pareadas de la validación cruzada de 10 folds externos. La pertinencia del test de Wilcoxon radica en que permite comparar el comportamiento de dos clasificadores evaluados bajo las mismas particiones de datos. En este caso, cada fold externo produjo una métrica de desempeño para BSC-BIO-EHR-ES y otra para el algoritmo clásico correspondiente. Al tratarse de observaciones emparejadas, la comparación no se realizó sobre promedios aislados, sino sobre las diferencias obtenidas fold por fold. Esta estructura permitió valorar si el modelo propuesto superaba de manera consistente a cada línea base o si las diferencias observadas podían explicarse por variaciones propias de las particiones de validación.

El procedimiento parte del cálculo de las diferencias entre los resultados de ambos modelos en cada fold. Posteriormente, se descartan las diferencias nulas, se ordenan las restantes según su valor absoluto y se asignan rangos. La prueba compara la suma de rangos asociados a diferencias positivas con la suma de rangos asociados a diferencias negativas. Si las diferencias entre modelos fueran aleatorias y no existiera una tendencia sistemática a favor de alguno de ellos, ambas sumas de rangos deberían ser relativamente similares. Por el contrario, una concentración marcada de rangos en una misma dirección sugiere que uno de los modelos presenta un rendimiento consistentemente superior.

$$W = \min(W^+, W^-) \quad (2.12)$$

2.7 Herramientas y plataformas tecnológicas

La implementación se apoyó en Google Colab Pro con GPU Tesla T4, mostrando que es posible obtener resultados competitivos dentro de una infraestructura accesible y sin recurrir a recursos de cómputo de escala industrial. Esta decisión fue metodológicamente relevante, ya que permitió entrenar y evaluar modelos de aprendizaje profundo bajo condiciones reproducibles, económicamente viables y compatibles con un entorno académico. Aunque el uso de infraestructura limitada impone restricciones en memoria, tiempo de ejecución y tamaño de lote, también permitió demostrar la factibilidad práctica del flujo experimental propuesto. Para el ajuste fino del modelo de lenguaje se utilizó Hugging Face Transformers, particularmente la clase Trainer y el optimizador AdamW, debido a que esta biblioteca proporciona herramientas estandarizadas para el entrenamiento, validación y evaluación de modelos basados en arquitecturas Transformer (Wolf et al., 2020). Su incorporación permitió organizar el proceso de entrenamiento de forma controlada, integrando particiones de datos, métricas de desempeño, criterios de parada y almacenamiento de resultados experimentales. Esto favoreció una implementación más reproducible y redujo la necesidad de construir manualmente cada componente del ciclo de entrenamiento.

El algoritmo genético fue implementado mediante DEAP, biblioteca orientada al diseño de procesos evolutivos y metaheurísticas de optimización. Esta herramienta permitió definir individuos, poblaciones, operadores de cruce, mutación y selección, adaptándolos al espacio vectorial de 768 dimensiones utilizado por las representaciones del modelo de lenguaje. De esta manera, DEAP funcionó como el soporte computacional para la generación de vectores sintéticos dentro del espacio latente, manteniendo consistencia con el diseño metodológico del aumento de datos. Por otra parte, scikit-learn se empleó para la implementación de los algoritmos clásicos de aprendizaje automático y para el cálculo de métricas de evaluación. Esta biblioteca permitió entrenar modelos supervisados y no supervisados bajo una misma estructura experimental, facilitando la comparación entre enfoques tradicionales y modelos de aprendizaje profundo. Asimismo, SciPy fue utilizado para aplicar las pruebas estadísticas de Wilcoxon y Friedman, necesarias para valorar si las diferencias observadas entre modelos podían considerarse estadísticamente significativas. Finalmente, Matplotlib se utilizó para la generación de visualizaciones, incluyendo representaciones t-SNE orientadas a explorar la distribución de los vectores en el espacio reducido.

Para la conversión de datos tabulares a lenguaje natural, correspondiente al flujo principal del proyecto, se utilizó la API de DeepSeek-R1 en modo de razonamiento avanzado (D. Guo et al., 2025a). Esta herramienta fue seleccionada por su viabilidad económica y por su capacidad para generar salidas razonadas que facilitaron la implementación de un protocolo de verificación supervisada de doble filtro. Dicho protocolo permitió revisar la coherencia entre los datos tabulares originales y las narrativas clínicas generadas, con el objetivo de reducir errores de omisión, contradicciones internas o inferencias no justificadas por los registros fuente. Este proceso transformó los 534 registros tabulares del Dataset balanceado, conformado por 267 casos positivos y 267 casos negativos, en historias clínicas narrativas estructuradas bajo los criterios de la NOM-004-SSA3-2012 (Secretaría de Salud, 2012). La conversión no tuvo como propósito inventar información clínica nueva, sino reorganizar los datos existentes en un formato textual compatible con modelos de lenguaje biomédico-clínico. Por ello, cada narrativa fue construida a partir de los atributos disponibles en el registro original, procurando mantener correspondencia semántica con las variables tabulares y evitando incorporar elementos no presentes en la fuente de datos.

La decisión de transformar los datos tabulares en narrativas clínicas respondió a la necesidad de alinear la estructura del Dataset con las capacidades del modelo BSC-BIO-EHR-ES. Dado que este tipo de modelo está diseñado para procesar lenguaje natural, la representación narrativa permitió aprovechar mejor su capacidad contextual frente a una entrada puramente tabular. En este sentido, el flujo principal de la investigación no consistió únicamente en clasificar registros, sino en evaluar si la conversión controlada de datos estructurados a texto clínico podía mejorar la representación semántica disponible para la tarea de clasificación binaria. No obstante, la generación de narrativas mediante un modelo de lenguaje fue tratada como una etapa metodológica sujeta a control y no como una fuente autónoma de verdad clínica. Las salidas generadas por DeepSeek-R1 fueron utilizadas como intermediarias textuales derivadas de datos tabulares previamente balanceados, por lo

que su validez dependió de la fidelidad con la información original y del proceso de verificación posterior. Esta distinción fue esencial para evitar que el modelo generador introdujera sesgos, ampliaciones injustificadas o contenido clínico no respaldado por los datos base.

En conjunto, la infraestructura y las herramientas empleadas permitieron construir un flujo experimental completo: generación narrativa de historiales clínicos, representación mediante modelo de lenguaje biomédico-clínico, aumento sintético en el espacio latente, entrenamiento de clasificadores, evaluación cruzada, validación estadística y visualización exploratoria. Esta integración metodológica permitió analizar el problema desde una perspectiva reproducible, combinando técnicas de procesamiento de lenguaje natural, aprendizaje profundo, aprendizaje automático clásico, optimización evolutiva y evaluación estadística.

CAPÍTULO III: ESTADO DEL ARTE

El presente capítulo ofrece una revisión sistemática de los trabajos más relevantes para esta investigación, organizada en cinco ejes temáticos: adaptación de modelos *Transformer* al dominio clínico, ajuste fino de *LLMs* con datos de salud limitados, *PLN* aplicado a registros clínicos electrónicos, sistemas conversacionales de salud basados en *IA*, y perspectivas complementarias sobre multilingüismo y ética. Para cada eje se desarrolla narrativamente el trabajo de mayor impacto, mientras que los trabajos complementarios se sintetizan en tablas especializadas. El análisis converge en la identificación de tres brechas: lingüística, de dominio oncológico y de validación estadística que esta investigación cubre simultáneamente.

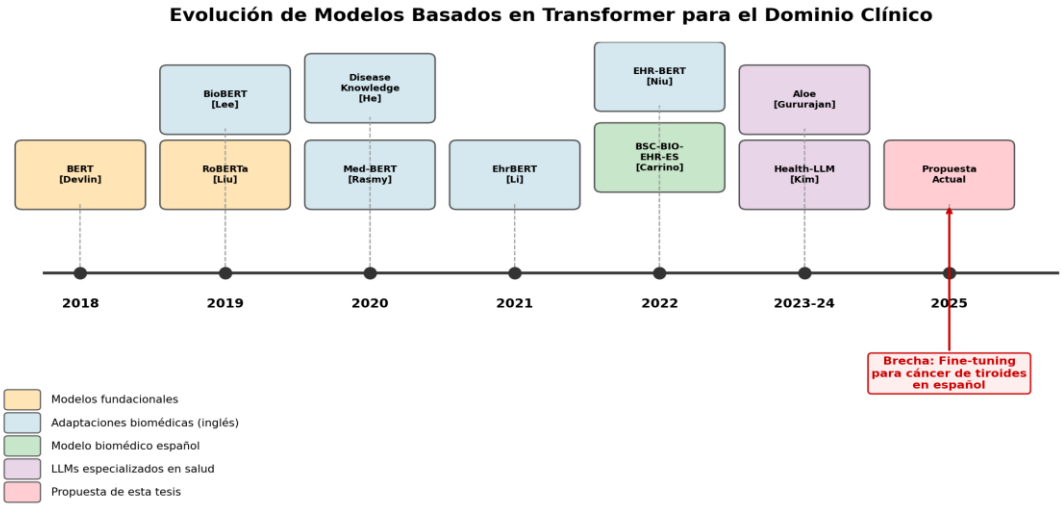

Figura 3.1: Evolución cronológica de los modelos basados en Transformer para el dominio clínico. Se observa la progresión desde los modelos fundacionales hasta la propuesta actual y la brecha que esta investigación aborda.

3.1 Adaptación de Modelos *BERT*/*Transformer* al Dominio Clínico

Esta línea de investigación constituye el antecedente más directo de la presente tesis. De los cinco trabajos revisados en este eje, se desarrolla a continuación el de (*Carrino* et al. 2022) por constituir el fundamento arquitectónico del modelo seleccionado; los cuatro restantes se muestran en la Tabla 3.1 y son desarrollados posteriormente.

***Pre-trained Biomedical Language Models for Clinical NLP in Spanish* ---** (Carrino et al., 2022)

En el marco del Plan de Tecnologías del Lenguaje del Gobierno de España, (*Carrino* et al. 2022) presentó los primeros modelos de lenguaje biomédicos entrenados desde cero para el español. Desarrollaron dos variantes basadas en la arquitectura *RoBERTa*: *bsc-bio-es*, entrenado exclusivamente con corpus biomédico general, y *bsc-bio-ehr-es*, que incorpora adicionalmente registros electrónicos de salud (*EHR*) reales. El objetivo fue crear modelos capaces de superar el rendimiento de alternativas generalistas y multilingües en tareas de *PLN* clínico en español.

Los modelos fueron entrenados con un corpus total de 1,100 millones de tokens limpios y deduplicados, incluyendo 95 millones provenientes de *EHR* reales. Se empleó la estrategia de preentrenamiento *RoBERTa* (*MLM* sin *NSP*) con un vocabulario *BPE* de 52,000 *tokens* optimizado para terminología médica en español. La evaluación abarcó tres tareas de *NER* clínico: *PharmaCoNER* (*F1: 0.8913*), *CANTEMIST* (*F1: 0.8340*) e *ICTUSnet* (*F1: 0.8756*). En todas, *bsc-bio-ehr-es* superó consistentemente a *mBERT*, *BETO-Galén*, *XLM-R-Galén* y *roberta-base-bne*. Este trabajo constituye el fundamento directo de la presente tesis, dado que *BSC-BIO-EHR-ES* es el modelo seleccionado para el proceso de ajuste fino, extendiéndolo por primera vez de *NER* a clasificación diagnóstica binaria en oncología tiroidea.

Tabla 3.1: Trabajos complementarios sobre adaptación de modelos BERT/Transformer al dominio clínico.

Artículo	Autor (Año)	Modelo	Objetivo	Idioma	Resultado Clave	Relevancia para esta Tesis
EHR-BERT: A BERT-based Model for Effective Anomaly Detection in EHR	*Niu et al. (2024)*	*EHR-BERT*	Detectar anomalías en HCE mediante Predicción Secuencial de Tokens Enmascarados (SMTP)	Inglés	*F1*: 0.80–0.99; reducción de falsos positivos	*BERT* adaptado detecta patrones anómalos clínicos; principio transferible a patrones oncológicos tiroideos
Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured EHR for Disease Prediction	*Rasmy et al. (2021)*	*Med-BERT*	Generar incrustaciones contextualizadas desde HCE de 28.4M pacientes con códigos CIE	Inglés	*AUC*: +1.21% a +6.14% en predicción de enfermedades	La ganancia del preentrenamiento específico es mayor con datos limitados análogo al régimen de 534 muestras
BEHRT: Transformer for Electronic Health Records	*Li et al. (2019)*	*EhrBERT*	Ajustar *BioBERT* con 1.5M notas HCE para normalización de entidades (MADE 1.0)	Inglés	*F1*: 0.4095; supera a *MetaMap* en 2.36%	Modelos en dominios cercanos superan a los lejanos valida selección de *BSC-BIO-EHR-ES*
Infusing Disease Knowledge into BERT for Health QA, Medical Inference and Disease Name Recognition	*He et al. (2020)*	*BERT+ Disease*	Infundir conocimiento de enfermedades en BERT mediante oraciones auxiliares	Inglés	Mejora global en QA, inferencia y NER en múltiples variantes BERT	Conversión tabular-a-lenguaje-natural bajo *NOM-004* es una infusión de conocimiento clínico mexicano

Desarrollo complementario de artículos sintetizados en la Tabla 3.1.

***EHR-BERT: A BERT-based Model for Effective Anomaly Detection in EHR* ---** (Niu et al., 2024)

Niu et al. (2024) propusieron *EHR-BERT*, un marco basado en la arquitectura *BERT* orientado a detectar anomalías en registros clínicos electrónicos mediante una estrategia de Predicción Secuencial de Tokens Enmascarados (*SMTP*), que trata las secuencias de eventos clínicos como oraciones de lenguaje natural y enmascara iterativamente sus tokens para aprender los patrones de ejecución normales en ambos sentidos. Evaluado sobre más de 16 millones de secuencias de *HCE* provenientes de tres dominios asistenciales, el modelo superó a los métodos previos de aprendizaje profundo, con valores de *F1* reportados entre 0.80 y 0.99 según el dominio, reduciendo de forma notable los falsos positivos y mejorando la tasa de detección (Niu et al., 2024). Aunque la tarea difiere de la de esta tesis, el principio de adaptar *BERT* para identificar patrones clínicos anómalos es transferible a la detección de patrones oncológicos tiroideos.

Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured EHR for Disease Prediction --- (Rasmy et al., 2021)

(Rasmy et al., 2021) adaptaron el marco *BERT* al dominio de las *HCE* estructuradas, preentrenando *Med-BERT* sobre los códigos diagnósticos (CIE-9 y CIE-10) de 28.49 millones de pacientes para generar incrustaciones contextualizadas reutilizables en tareas de predicción de enfermedades. El ajuste fino de estas incrustaciones incrementó el área bajo la curva (*AUC*) entre 1.21% y 6.14% respecto a modelos sin preentrenamiento y, de manera relevante para esta investigación, la ganancia fue mayor en escenarios con conjuntos de entrenamiento pequeños, donde *Med-BERT* alcanzó un *AUC* comparable al de modelos entrenados con diez veces más datos (Rasmy et al., 2021). Este comportamiento respalda que el preentrenamiento específico de dominio es especialmente ventajoso en regímenes de escasez de datos, análogo al de las 534 muestras de esta tesis.

***Fine-Tuning BERT-Based Models on Large-Scale Electronic Health Record Notes* ---** (Li et al., 2019)

Li et al. (2019) investigaron el efecto del dominio de los datos de entrenamiento sobre el rendimiento de modelos basados en *BERT* en la normalización de entidades clínicas. Para ello ajustaron *BioBERT* sobre 1.5 millones de notas no etiquetadas de *HCE* generando el modelo *EhrBERT* y lo evaluaron en los corpus *MADE 1.0*, *NCBI Disease* y *CDR*. *EhrBERT* alcanzó un *F1* de 0.4095 en *MADE 1.0*, superando a *MetaMap* en 2.36 puntos porcentuales, y mejoró a *DNorm* en los corpus de enfermedades (Li et al., 2019). El hallazgo de que la proximidad entre el dominio de preentrenamiento y la tarea objetivo eleva el rendimiento valida la selección de un modelo de dominio cercano como *BSC-BIO-EHR-ES* en esta tesis.

***Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition* ---** (He et al., 2020)

He et al. (2020) propusieron un procedimiento de infusión de conocimiento sobre enfermedades que enriquece modelos tipo *BERT* a partir de artículos estructurados de enfermedades, entrenando al modelo para inferir el aspecto de la enfermedad asociado a cada fragmento de texto. Aplicado a múltiples variantes (entre ellas *BERT*, *BioBERT* y *BlueBERT*), el método mejoró el rendimiento en casi todos los casos en tres tareas respuesta a preguntas de salud, inferencia médica y reconocimiento de nombres de enfermedades; por ejemplo, la exactitud de *BioBERT* en respuesta a preguntas de salud del consumidor aumentó de 68.29% a 72.09%, estableciendo nuevos resultados de vanguardia en dos conjuntos de datos (He et al., 2020). Por analogía, la conversión de datos tabulares a lenguaje natural bajo la *NOM-004* puede interpretarse como una forma de infusión de conocimiento clínico contextualizado al entorno mexicano.

Síntesis. El patrón dominante es claro: la adaptación de Transformers al dominio clínico produce mejoras consistentes, amplificadas cuando existe proximidad entre dominio de preentrenamiento y tarea objetivo. La presente tesis extiende este paradigma aplicando *BSC-BIO-EHR-ES* a clasificación diagnóstica oncológica.

3.2 Ajuste Fino de *LLMs* con Datos de Salud Limitados

El *Dataset* de 534 muestras y una ratio parámetros/muestra de $\sim 2\times10^{5}$:1 sitúan este proyecto en un régimen de escasez de datos. De los cuatro trabajos revisados, se desarrolla (Xu et al., 2024) por aportar la evidencia más directa sobre viabilidad del ajuste fino con pocas muestras.

***Mental-LLM: Leveraging Large Language Models for Mental Health Prediction* ---** (Xu et al., 2024)

(*Xu* et al. 2024) realizaron una evaluación exhaustiva de múltiples *LLMs* (*Alpaca*, *FLAN-T5*, *GPT-3.5*, *LLaMA2* y *GPT-4*) en tareas de predicción de salud mental utilizando datos de texto en línea. Se compararon sistemáticamente tres enfoques: *zero-shot*, *few-shot* e *instruction fine-tuning*, evaluando el impacto del volumen y la diversidad de los datos de ajuste fino sobre múltiples tareas y conjuntos de datos.

El ajuste fino de instrucciones incrementó significativamente el rendimiento para todas las tareas de manera simultánea. Dos hallazgos resultan determinantes para esta tesis: primero, que el ajuste fino con solo unos pocos cientos de muestras de múltiples conjuntos de datos ya logra un rendimiento favorable; y segundo, que la diversidad de los datos es más determinante que su volumen. Esta evidencia respalda directamente la viabilidad del ajuste fino de *BSC-BIO-EHR-ES* con 534 muestras, donde la riqueza semántica de las narrativas clínicas generadas bajo la *NOM-004* compensa la limitación cuantitativa del corpus.

Tabla 3.2: Trabajos complementarios sobre ajuste fino de LLMs con datos de salud limitados.

Artículo	Autor (Año)	Modelo Base	Técnica FT	Dataset	Resultado Clave	Relevancia para esta Tesis
Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data	*Kim et al. (2024)*	*HealthAlpaca*	*LoRA* + enriquecimiento de contexto	Multi-tarea	Supera a *GPT-3.5/4* en 8/10 tareas; +23.8%	Modelo pequeño con *FT* supera a modelos masivos en tareas específicas
Aloe: A Family of Fine-tuned Open Healthcare LLMs	*Gururajan et al. (2024)*	*Mistral / LLaMA 3*	*DPO* + datos sintéticos con *CoT*	Sintético + público	Nuevo estándar para *LLMs* médicos abiertos y alineados	Generación de datos sintéticos con *CoT* relevante para conversión tabular vía *DeepSeek R1 SOTA*
HealAI: LLM para Documentación Médica	*Goyal et al. (2024)*	*LLM custom*	*SFT + RAG* para conversaciones largas	Documentación clínica	Supera a *GPT-4* en redacción de notas médicas	Confirma superioridad del ajuste fino específico frente a *LLMs* masivos

Desarrollo complementario de artículos sintetizados en la Tabla 3.2.

***Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data* ---** (Kim et al., 2024)

Kim et al. (2024) evaluaron doce *LLMs* de vanguardia, mediante prompting y ajuste fino, en diez tareas de predicción de salud a partir de datos de sensores vestibles. Su modelo ajustado *HealthAlpaca* de una arquitectura comparativamente pequeña igualó o superó a modelos mucho mayores como *GPT-3.5*, *GPT-4* y *Gemini-Pro*, obteniendo el mejor desempeño en 8 de las 10 tareas, mientras que las estrategias de enriquecimiento de contexto aportaron mejoras de hasta 23.8% (Kim et al., 2024). Este resultado evidencia que un modelo de menor tamaño, ajustado a la tarea, puede aventajar a modelos masivos generalistas, lo que respalda el enfoque de ajuste fino adoptado en esta tesis.

***Aloe: A Family of Fine-tuned Open Healthcare LLMs* ---** (Gururajan et al., 2024)

Gururajan et al. (2024) presentaron la familia *Aloe* de modelos médicos abiertos, construidos a partir de modelos base como *Mistral* y *LLaMA 3* y entrenados con un conjunto de datos que combina fuentes públicas enriquecidas con datos sintéticos generados mediante Cadena de Pensamiento (*CoT*). Tras una fase de alineación con Optimización Directa de Preferencias (*DPO*) y el uso de estrategias avanzadas de inferencia, *Aloe* se posicionó como uno de los primeros *LLMs* de salud abiertos y alineados éticamente, competitivo dentro de su rango de escala (Gururajan et al., 2024). El empleo de datos sintéticos de calidad guiados por *CoT* es directamente pertinente para la conversión tabular-a-lenguaje-natural realizada en esta tesis.

***HealAI: A Healthcare LLM for Effective Medical Documentation* ---** (Goyal et al., 2024)

Goyal et al. (2024) describieron el desarrollo de un *LLM* médico costo-eficiente especializado en la redacción de notas clínicas, optimizado mediante ajuste fino supervisado (*SFT*) frente a las limitaciones de la ingeniería de prompts sobre modelos genéricos. A través de comparaciones sistemáticas de tamaños de modelo, datos de entrenamiento y diseños de tubería, el modelo superó a *GPT-4* en tareas de documentación médica empleando modelos más pequeños, manteniendo la precisión y reduciendo sesgos y alucinaciones (Goyal et al., 2024). Este trabajo confirma la superioridad del ajuste fino específico sobre los *LLMs* masivos generalistas, premisa central de esta investigación.

Síntesis. La evidencia converge en tres principios: modelos pequeños con ajuste fino superan a modelos masivos generalistas, la diversidad importa más que el volumen, y los datos sintéticos de calidad son viables en régimen de escasez. Estos hallazgos sustentan la viabilidad técnica del ajuste fino con 534 muestras.

3.3 PLN Aplicado a Registros Clínicos Electrónicos

Esta línea de investigación valida la aplicación de modelos *Transformer* sobre texto libre de *HCE*. Se desarrolla Lee et al. (2023) por constituir el paralelo metodológico más directo con la presente tesis.

Assessment of NLP of Electronic Health Records as a Clinical Trial Outcome --- (R. Y. Lee et al., 2023)

Lee et al. (2023) evaluaron el uso de *PLN* basado en aprendizaje profundo para medir discusiones sobre objetivos de atención documentadas en *HCE*, en el contexto de un ensayo clínico pragmático aleatorizado. Se compararon tres enfoques: *PLN* de aprendizaje profundo (*Bio+ClinicalBERT*), abstracción humana cribada por PLN, y abstracción manual convencional.

Bio+ClinicalBERT, entrenado con un conjunto de datos separado, alcanzó un F1 de 0.82, un *ROC-AUC* de 0.924 y un *PR-AUC* de 0.879 en una muestra de validación de 159 participantes del ensayo. Este trabajo demuestra la viabilidad de aplicar modelos *BERT* clínicos para extraer información compleja del texto libre de las *HCE*, con rendimiento comparable al de abstracción humana. El paralelismo con esta tesis es directo: ambos proyectos emplean un modelo *BERT* clínico sobre texto libre para una tarea de clasificación binaria, con validación en muestras clínicas reales.

Tabla 3.3: Trabajos complementarios sobre PLN aplicado a registros clínicos electrónicos.

Artículo	Autor (Año)	Arquitectura	Tipo HCE	Tarea	Resultado Clave	Relevancia para esta Tesis
Named Entity Recognition Method in Health Preserving Field Based on BERT	Zhang et al. (2021)	*BERT-CNN-BiLSTM-CRF*	Dominio preservación de salud	NER para grafos de conocimiento	Alto rendimiento y estabilidad desde primeras épocas	Evidencia la capacidad de *BERT* para identificar entidades cruciales en textos médicos
EMR-Driven ML Predictive Model for Hospital-Acquired Pressure Injuries	Nguyen et al. (2025)	*XGBoost* + eliminación recursiva	Datos longitudinales HCE	Predicción de lesiones por presión	AUROC: 0.83–0.85; validación en 5 hospitales; 98% precisión etiquetada	Pipeline de curación y validación multicéntrica como referencia metodológica

Desarrollo complementario de artículos sintetizados en la Tabla 3.3.

***Named Entity Recognition Method in Health Preserving Field Based on BERT* ---** (Zhang et al., 2021)

Zhang et al. (2021) propusieron un método de reconocimiento de entidades nombradas (*NER*) para el dominio de la preservación de la salud que, ante la ausencia de conjuntos de datos públicos, construye un corpus a partir de fuentes web y define siete tipos de entidades. La arquitectura combina *BERT* que genera representaciones contextuales del mismo término según su contexto con una capa *CNN* para extraer rasgos locales, una *BiLSTM* para capturar dependencias de larga distancia y un campo aleatorio condicional (*CRF*) para decodificar la mejor secuencia de etiquetas, alcanzando alto rendimiento y estabilidad desde las primeras épocas de entrenamiento (Zhang et al., 2021). Esta evidencia confirma la capacidad de *BERT* para identificar entidades clínicas cruciales en texto médico, capacidad que esta tesis aprovecha para la clasificación diagnóstica.

***Electronic-Medical-Record-Driven Machine Learning Predictive Model for Hospital-Acquired Pressure Injuries* ---** (Nguyen et al., 2025)

Nguyen et al. (2025) desarrollaron y validaron externamente un modelo de aprendizaje automático para predecir el riesgo de lesiones por presión adquiridas en el hospital a partir de datos longitudinales de *HCE*. El sistema empleó *XGBoost* con eliminación recursiva de características para seleccionar 35 variables clínicas óptimas y un análisis de series temporales para la predicción dinámica del riesgo, sustentado en una tubería automatizada de curación, etiquetado e integración de datos; la validación interna y multicéntrica sobre 5510 hospitalizaciones en cinco hospitales arrojó valores de *AUROC* de 0.83–0.85, superando a la escala de Braden (Nguyen et al., 2025). Su tubería de curación y su esquema de validación multicéntrica constituyen una referencia metodológica para el diseño experimental de esta tesis.

Síntesis. Los trabajos demuestran que el *PLN* sobre *HCE* extrae información clínica compleja con rendimiento comparable al de expertos humanos. La combinación de arquitecturas *Transformer* con protocolos rigurosos de validación establece estándares adoptados en el diseño experimental de esta tesis.

3.4 Agentes Conversacionales y Sistemas Interactivos en Salud

Si bien la presente tesis no implementa agentes conversacionales, esta línea aporta evidencia sobre técnicas de razonamiento estructurado y estrategias de composición de datos que fueron adoptadas en el flujo metodológico. Se desarrolla Alghamdi y Mostafa (2024) por su evidencia cuantitativa sobre el impacto de la proporción de datos en el rendimiento de modelos clínicos.

***Towards Reliable Healthcare LLM Agents* ---** (Alghamdi & Mostafa, 2024)

Alghamdi y Mostafa (2024) presentaron una metodología integral para mejorar la confiabilidad de agentes LLM en salud, con aplicación a la atención de peregrinos durante el *Hajj*. Crearon el dataset *HajjHealthQA* combinando datos de fuentes oficiales (Ministerio de Salud de Arabia Saudita, *OMS*) con datos generados, y realizaron ajuste fino sobre *GPT-3.5 Turbo* y *Llama3* comparando diferentes proporciones en la composición del corpus de entrenamiento.

El hallazgo más relevante para esta tesis es que la estrategia de segmentación *50/50* equilibrando datos de origen real con datos generados alcanzó un rendimiento de *83.5%*, demostrando que una composición balanceada del corpus de entrenamiento produce modelos robustos sin depender exclusivamente de una sola fuente de datos. Este principio de equilibrio en la composición del corpus es directamente análogo al enfoque de esta tesis, donde la conversión tabular-a-lenguaje-natural mediante la *API* de *DeepSeek R1* en modo *SOTA* emplea Cadena de Pensamiento (*CoT*) para producir narrativas clínicas que preservan la integridad del dato clínico original dentro de un corpus balanceado de 534 muestras (267 positivas, 267 negativas).

Tabla 3.4: Trabajos complementarios sobre agentes conversacionales y sistemas interactivos en salud.

Artículo	Autor (Año)	Sistema	Técnicas	Resultado Clave	Relevancia para esta Tesis
HealthQ: Unveiling Questioning Capabilities of LLM Chains in Healthcare Conversations	Wang et al. (2025)	*HealthQ*	*CoT*, *ReAct*, cadenas reflexivas	Cadenas con razonamiento estructurado superan a las simples en obtención de información clínica	Valida *CoT* como técnica de razonamiento clínico; proyecta extensión futura del modelo
Conversational Health Agents: A Personalized LLM-Powered Agent Framework	Abbasian et al. (2024)	*openCHA*	*Tree of Thought*, *ReAct*, Fuentes externas	Respuestas personalizadas integrando guías clínicas externas	Integración de guías *ATA* es extensión natural del modelo
BianQue: Balancing Questioning and Suggestion Ability of Health LLMs	Chen et al. (2023)	*BianQue*	*CoQ*; *2.4M* muestras pulidas con *ChatGPT*	Equilibrio preguntas/sugerencias; 46.2% respuestas son preguntas	Diagnóstico tiroideo requiere recopilación progresiva multivuelta

Desarrollo complementario de artículos sintetizados en la Tabla 3.4.

***HealthQ: Unveiling Questioning Capabilities of LLM Chains in Healthcare Conversations* ---** (Wang et al., 2025)

Wang et al. (2025) propusieron *HealthQ*, un marco para evaluar la capacidad de las cadenas de *LLM* de formular preguntas que recaben información clínica relevante. Implementaron y compararon arquitecturas de cadena entre ellas generación aumentada por recuperación (*RAG*), Cadena de Pensamiento (*CoT*) y cadenas reflexivas y emplearon un *LLM* evaluador junto con métricas tradicionales de *PLN* (*ROUGE* y comparación basada en *NER*) sobre conjuntos derivados de *ChatDoctor* y *MTS-Dialog*. Los resultados mostraron que las cadenas dotadas de razonamiento estructurado superan a las simples en la obtención de información clínica (Wang et al., 2025). Este hallazgo valida el uso de *CoT* como técnica de razonamiento clínico y proyecta una extensión futura del modelo propuesto.

***Conversational Health Agents: A Personalized LLM-Powered Agent Framework* ---** (Abbasian et al., 2024)

Abbasian et al. (2024) propusieron *openCHA*, un marco de código abierto basado en *LLM* que dota a los agentes conversacionales de salud de capacidades de las que carecían los sistemas previos: resolución de problemas en múltiples pasos, conversación personalizada y análisis multimodal. El marco incorpora un orquestador que planifica y ejecuta acciones para recuperar información de fuentes externas datos, bases de conocimiento y modelos de análisis y la integra para generar respuestas personalizadas a las consultas de salud (Abbasian et al., 2024). La integración de guías clínicas externas que habilita este marco representa una extensión natural del modelo de esta tesis, por ejemplo, hacia las directrices de la *ATA*.

***BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs* ---** (Chen et al., 2023)

Chen et al. (2023) abordaron la limitación de los *LLMs* de salud para conducir interrogatorios de múltiples turnos, formalizando la noción de cadena de preguntas (*CoQ*). Para ello ajustaron un modelo basado en *ChatGLM* con el corpus autoconstruido *BianQueCorpus* compuesto por cerca de 2.4 millones de conversaciones de salud multivuelta pulidas con *ChatGPT*, en el que las preguntas representan el 46.2% de las respuestas del médico, logrando equilibrar la capacidad de preguntar con la de sugerir (Chen et al., 2023). Dado que el diagnóstico tiroideo requiere una recopilación progresiva de información a lo largo de varios turnos, este enfoque resulta pertinente para futuras extensiones interactivas del modelo.

Síntesis. La evidencia de esta sección converge en dos ejes adoptados por esta tesis: el razonamiento estructurado (*CoT*) como mecanismo para preservar la coherencia clínica en la generación de narrativas, y el equilibrio en la composición del corpus como factor determinante del rendimiento. Los agentes conversacionales como tales proyectan una extensión futura del modelo hacia sistemas de apoyo al diagnóstico interactivos.

3.5 Perspectivas Complementarias: *Reviews*, Multilingüismo y Ética

Esta sección reúne trabajos que proporcionan fundamentos teóricos, evidencia empírica sobre multilingüismo y consideraciones éticas. Se desarrolla Skianis et al. (2025) por constituir la evidencia empírica central de la brecha lingüística.

Building Multilingual Datasets for Predicting Mental Health Severity through LLMs: Prospects and Challenges --- (Skianis et al., 2024)

Este estudio presenta la creación de un nuevo conjunto de datos multilingüe diseñado para evaluar cómo los modelos de lenguaje extenso (LLM) detectan la gravedad de problemas de salud mental. Los autores tradujeron publicaciones de redes sociales del inglés a seis idiomas, incluyendo griego, turco y alemán, para analizar el desempeño de modelos guiados a un enfoque de salud. Los resultados revelan una variabilidad significativa en la precisión según el idioma, lo que resalta las dificultades de captar matices lingüísticos en contextos médicos. Se destaca que, aunque este enfoque es económicamente eficiente para expandir recursos en idiomas de bajo nivel, existen riesgos de diagnósticos erróneos si se usan sin supervisión humana.

Tabla 3.5: Perspectivas complementarias reviews, multilingüismo y consideraciones éticas.

Artículo	Autor (Año)	Tipo	Tema Central	Hallazgo / Conclusión Clave	Relevancia para esta Tesis
Large Language Models in Biomedical and Health Informatics: A Review	(Yu et al., 2024)	*Review*	*LLMs* biomédicos y clínicos	Modelos generales tienen eficacia limitada en tareas especializadas; modelos de dominio logran rendimiento superior	Valida la estrategia de emplear *BSC-BIO-EHR-ES* frente a alternativas generalistas
LLMs in Medical Specialties: A Comprehensive Review	(Mumtaz et al., 2023)	*Review*	*LLMs* en especialidades médicas	Los *LLMs* son *few-shot learners*: adaptables con pocos ejemplos específicos	Respalda la viabilidad del ajuste fino con 534 muestras
Opportunities and Risks of LLMs in Mental Health	Lawrence et al. (2024)	Análisis	*LLMs* en salud mental	Modelos ajustados con datos curados mitigan riesgos de modelos generales	Justifica el ajuste fino supervisado con datos clínicos curados
BERT: A Review of Applications in NLP and Understanding	(Koroteev, 2021)	*Review técnico*	Arquitectura y adaptación de *BERT*	Preentrenamiento continuado con tasas de aprendizaje bajas previene olvido catastrófico	Fundamenta directamente la $lr = 2\times10^{-5}$ empleada
The Impact of Multimodal LLMs on Health Care	Meskó (2023)	Prospectivo	*LLMs* multimodales en salud	Proyecta interpretación simultánea de texto, imágenes y datos genómicos	Posiciona la extensión multimodal como trabajo futuro natural
LLMs in Health Professions Education Research	Ellaway y Tolsgaard (2023)	Ético	*LLMs* en investigación académica	Necesidad de transparencia en uso de IA generativa en investigación	Sustenta documentación explícita de *DeepSeek R1* en el flujo metodológico

Desarrollo complementario de artículos sintetizados en la Tabla 3.5.

Large Language Models in Biomedical and Health Informatics: A Review **---** (Yu et al., 2024)

(Yu et al., 2024) ofrecieron una revisión con análisis bibliométrico de las aplicaciones de los *LLMs* en la informática biomédica y de la salud, sistematizando un amplio conjunto de artículos por temáticas y categorías diagnósticas y mapeando las redes de colaboración del campo. Entre sus conclusiones destaca que los modelos de propósito general presentan una eficacia limitada en tareas especializadas, mientras que los modelos de dominio alcanzan un rendimiento superior (Yu et al., 2024). Esta observación valida la estrategia de esta tesis de emplear un modelo de dominio como *BSC-BIO-EHR-ES* frente a alternativas generalistas.

LLMs in Medical Specialties: A Comprehensive Review **---** (Mumtaz et al., 2023)

(Mumtaz et al., 2023) presentaron una revisión de las aplicaciones y desafíos de los *LLMs* en distintas especialidades médicas incluyendo el cuidado del cáncer, la dermatología, la odontología, los trastornos neurodegenerativos y la salud mental, con énfasis en sus funcionalidades diagnósticas y terapéuticas. La revisión reconoce la capacidad de los *LLMs* para adaptarse a tareas clínicas específicas a partir de pocos ejemplos, en línea con su naturaleza de *few-shot learners* (Mumtaz et al., 2023). Esta propiedad respalda la viabilidad del ajuste fino con un corpus limitado de 534 muestras.

Opportunities and Risks of LLMs in Mental Health **---** (Lawrence et al., 2024)

Lawrence et al. (2024) revisaron la literatura sobre el uso de *LLMs* en educación, evaluación e intervención en salud mental, identificando oportunidades de impacto positivo junto con riesgos asociados a su despliegue. Entre sus recomendaciones centrales destaca la necesidad de que los *LLMs* destinados a la salud mental sean ajustados específicamente al dominio con datos curados, además de atender la equidad y los estándares éticos, como vía para mitigar los riesgos de los modelos de propósito general (Lawrence et al., 2024). Este argumento justifica el ajuste fino supervisado con datos clínicos curados adoptado en esta tesis.

BERT: A Review of Applications in NLP and Understanding **---** (Koroteev, 2021)

(Koroteev, 2021) sistematizó el mecanismo de funcionamiento de *BERT* y sus principales áreas de aplicación en el análisis de texto, comparándolo con modelos afines y describiendo variantes de dominio. La revisión documenta cómo el ajuste fino y el preentrenamiento continuado, con procedimientos de entrenamiento cuidadosamente parametrizados, resultan determinantes para el desempeño de los modelos derivados de *BERT* (Koroteev, 2021). Estas consideraciones fundamentan decisiones de configuración del ajuste fino en esta tesis, como el empleo de tasas de aprendizaje bajas ($lr = 2\times10^{-5}$) para preservar el conocimiento preentrenado y evitar el olvido catastrófico.

The Impact of Multimodal LLMs on Health Care **---** (Meskó, 2023)

Meskó (2023) analizó, desde una perspectiva prospectiva, el impacto de los *LLMs* multimodales (*M-LLMs*) en el futuro de la atención sanitaria, es decir, modelos capaces de interpretar y generar no solo texto sino también imágenes, video y voz. El autor plantea escenarios en los que los *M-LLMs* actúan como puerta de enlace entre los profesionales de la salud y la inteligencia artificial, integrando texto, imágenes e incluso datos genómicos, sin sustituir el componente humano de la medicina (Meskó, 2023). Esta visión posiciona la extensión multimodal del modelo como una línea de trabajo futuro natural para esta tesis.

LLMs in Health Professions Education Research **---** (R. H. Ellaway & Tolsgaard, 2023)

Ellaway y Tolsgaard (2023), en un editorial sobre la investigación en educación de las profesiones de la salud, examinaron las implicaciones de los *LLMs* como *ChatGPT* para la autoría y la autoridad de los trabajos académicos, así como los retos éticos asociados. Los autores enfatizan la necesidad de transparencia en el uso de la inteligencia artificial generativa en la investigación (Ellaway y Tolsgaard, 2023). Este principio sustenta la documentación explícita del uso de *DeepSeek R1* en el flujo metodológico de esta tesis.

Síntesis. Las perspectivas complementarias refuerzan tres decisiones fundamentales de esta tesis: la selección de un modelo nativo en español (por la variabilidad interlingüística demostrada), el uso de ajuste fino supervisado con datos curados (frente a los riesgos documentados de modelos generales), y la transparencia metodológica en el uso de herramientas de *IA* generativa.

3.6 Análisis de Brechas y Posicionamiento de la Propuesta

El análisis de los trabajos revisados permite identificar tres brechas fundamentales que la presente investigación cubre simultáneamente.

Brecha lingüística. La totalidad de los trabajos con ajuste fino clínico operan en inglés. Skianis et al. (2025) demostraron que el rendimiento varía significativamente entre idiomas, invalidando la traducción como estrategia. Carrino et al. (2022) probaron la viabilidad de modelos nativos en español, pero limitados a NER. Esta tesis extiende ese esfuerzo al ajuste fino para clasificación diagnóstica oncológica.

Brecha de dominio oncológico. Ningún trabajo aborda el diagnóstico de cáncer de tiroides mediante modelos de lenguaje. Los dominios más cercanos son anomalías genéricas en *EHR* (Niu et al., 2024), enfermedades crónicas (Rasmy et al., 2021) y salud mental (Xu et al., 2024). Esta tesis constituye uno de los primeros esfuerzos en aplicar modelos de lenguaje biomédicos en español al diagnóstico de cáncer tiroideo.

Brecha de validación estadística. La mayoría de trabajos emplea *hold-out* simple o *k-fold* básico, sin pruebas no paramétricas formales. El *Nested Cross-Validation* (10×3) con pruebas de *Friedman y Wilcoxon* (*p-values* $\sim 10^{-15}$) de esta tesis proporciona un nivel de certeza que supera ampliamente los estándares reportados.

La Tabla 3.6 muestra la evidencia marcada por brecha, su evidencia, la cobertura de la propuesta y el articulo que la respaldaway

Tabla 3.6: Síntesis cruzada de brechas identificadas y posicionamiento de la propuesta.

Brecha	Evidencia	Cobertura por la Propuesta	Artículos de Respaldo
Lingüística	*100%* de trabajos con *FT* clínico operan en inglés; rendimiento varía entre idiomas (Skianis, 2025)	Ajuste fino de *BSC-BIO-EHR-ES*, modelo nativo en español biomédico con *EHR* reales	Carrino (2022), Skianis (2025), (Yu et al., 2024)
Dominio oncológico	Ningún trabajo aborda cáncer de tiroides con modelos de lenguaje; dominios cubiertos: anomalías *HCE*, enf. crónicas, salud mental	Clasificación diagnóstica binaria de cáncer tiroideo con narrativas clínicas bajo *NOM-004*	Niu (2024), Rasmy (2021), Xu (2024), Lee (2023)
Validación estadística	Mayoría usa *hold-out* simple o *k-fold* básico; ausencia de pruebas no paramétricas formales	*Nested CV* (10×3) + *Friedman* + *Wilcoxon* ($p < 10^{-15}$) vs. 6 clasificadores clásicos	Todos los revisados (por contraste)

En conjunto, estas tres brechas posicionan la presente investigación como una contribución original al campo de la inteligencia artificial médica en el contexto hispanohablante, demostrando que resultados competitivos son alcanzables con recursos limitados cuando se aplica una metodología rigurosa y se aprovechan los avances en modelos de lenguaje preentrenados.

CAPÍTULO IV: METODOLOGÍA DE SOLUCIÓN

El presente capítulo describe el protocolo metodológico diseñado para desarrollar y evaluar un sistema de soporte diagnóstico para cáncer de tiroides en español, mediante el ajuste fino del modelo de lenguaje *BSC-BIO-EHR-ES*. La exposición integra los objetivos de la investigación en la lógica narrativa de cada fase, de modo que el vínculo entre el qué se hace y el para qué se hace sea explícito a lo largo del texto. El capítulo establece decisiones de diseño, criterios de selección y procedimientos operacionales reproducibles.

4.1 Diseño general de la investigación

4.1.1 Enfoque, alcance y tipo de estudio

La investigación adopta un enfoque cuantitativo-experimental inscrito en el paradigma positivista-computacional. La variable de interés central el desempeño diagnóstico del modelo de lenguaje se cuantifica mediante indicadores numéricos objetivos (*F1-Score, ROC-AUC, ECE, Brier Score*), cuya variación responde a la manipulación controlada de la estrategia de entrenamiento. La formulación de hipótesis a priori, la operacionalización explícita de variables y el uso de pruebas estadísticas formales configuran el estudio como investigación experimental de tipo aplicado (Devlin et al., 2019).

El alcance es explicativo-confirmatorio: el estudio somete a prueba predicciones cuantitativas sobre el efecto de la estrategia de ajuste fino en el rendimiento diagnóstico bajo condiciones de escasez de datos. Con 125 millones de parámetros ajustándose sobre 534 muestras, la relación parámetro-muestra del presente trabajo representa uno de los regímenes de mayor riesgo de sobreajuste documentados en la literatura de ajuste fino de Transformers (Mosbach et al., 2021a), lo que hace indispensable la justificación metodológica detallada de cada decisión de diseño. El estudio se enmarca como investigación aplicada de desarrollo tecnológico de base, no como sistema de despliegue clínico inmediato.

4.1.2 Correspondencia entre problema, objetivos, variables e indicadores

La coherencia interna del diseño metodológico se sustenta en la articulación explícita entre el problema central de investigación, los objetivos planteados, las variables de análisis y los indicadores de evaluación. Esta correspondencia permite asegurar que cada fase del estudio responde a una necesidad científica concreta y que los procedimientos aplicados generan evidencia pertinente para el contraste posterior de las hipótesis, cuya formulación formal se presenta en la *sección 4.5.3.1*. El objetivo general, orientado a identificar características de la sintomatología del cáncer de tiroides para apoyar la generación de un diagnóstico con información médica en español, estructura el conjunto del flujo metodológico. Su evaluación se apoya en indicadores de desempeño diagnóstico, principalmente *F1-Score* y *ROC-AUC*, complementados por métricas de calibración y estabilidad. En términos metodológicos, este objetivo exige comparar el rendimiento del modelo propuesto frente a enfoques alternativos que operan sobre la misma base de datos, pero mediante representaciones distintas.

El *OE1*, referido a la observancia de la normatividad ética y legal en la compilación de corpus clínicos en español, orienta la *Fase I* y la *Fase III*. En la primera, este objetivo se concreta en la selección, depuración, anonimización y resguardo del *Dataset* clínico tabular; en la tercera, en la construcción de narrativas clínicas bajo el formato de la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012). Los indicadores asociados a este objetivo son la integridad del corpus, la consistencia documental del pipeline de conversión y el cumplimiento de criterios de resguardo ético y técnico.

El *OE2*, orientado a detectar patrones o particularidades del *Dataset* que puedan afectar el comportamiento del sistema, se desarrolla en la *Fase II* mediante la evaluación de la representación tabular original. Esta fase incluye tanto la exploración preliminar de separabilidad natural como la construcción de una línea base supervisada con algoritmos clásicos. Las variables asociadas corresponden al comportamiento discriminativo del conjunto de datos bajo representación tabular, y sus indicadores principales son *F1-Score*, *ROC-AUC*, *Accuracy*, *Precisión* y *Recall*, además de métricas exploratorias de agrupamiento en los experimentos no supervisados.

El *OE3*, enfocado en optimizar la capacidad del sistema para producir inferencias diagnósticas en español, se articula principalmente con la *Fase IV*. En esta fase se integran la selección del modelo base, el ajuste fino supervisado y la estrategia de optimización bajo escasez de datos mediante aumento sintético en espacio latente. Las variables de interés incluyen el desempeño del modelo, su estabilidad entre particiones, su convergencia durante el entrenamiento y su comportamiento probabilístico. Los indicadores correspondientes son *F1-Score*, desviación estándar del *F1* entre pliegues, curvas de pérdida de entrenamiento y validación, *ECE* y *Brier Score*.

El *OE4*, orientado a la comparación con métodos de referencia, estructura transversalmente la *Fase II* y la *sección 4.6*. Su función es establecer una base objetiva de contraste entre el modelo propuesto, los algoritmos clásicos que operan sobre el *Dataset* tabular y los modelos de referencia en configuración *Zero-shot*. Este objetivo se evalúa mediante comparaciones de desempeño por pliegue y pruebas estadísticas no paramétricas, específicamente *Wilcoxon* y *Friedman*, aplicadas sobre los vectores de métricas obtenidos bajo validación cruzada anidada.

Finalmente, el *OE5*, dirigido a identificar fortalezas, limitaciones y áreas de mejora del sistema, se vincula con el análisis de calibración, interpretabilidad y comparación entre variantes del modelo desarrolladas en la *Fase IV*. Este objetivo incorpora indicadores de calidad probabilística, robustez semántica y criterios de selección entre la variante base y la variante optimizada, con el fin de determinar cuál representa de forma más adecuada el artefacto final de la investigación.

4.1.3 Flujo metodológico general

El flujo metodológico del estudio comprende cuatro fases que operan sobre representaciones progresivamente más ricas de la información clínica: el dato tabular original, su evaluación como representación directa para clasificación, su conversión a lenguaje natural, y el desarrollo y optimización del modelo de lenguaje sobre esa representación textual. *La figura 4.1* ilustra este flujo con sus transiciones.

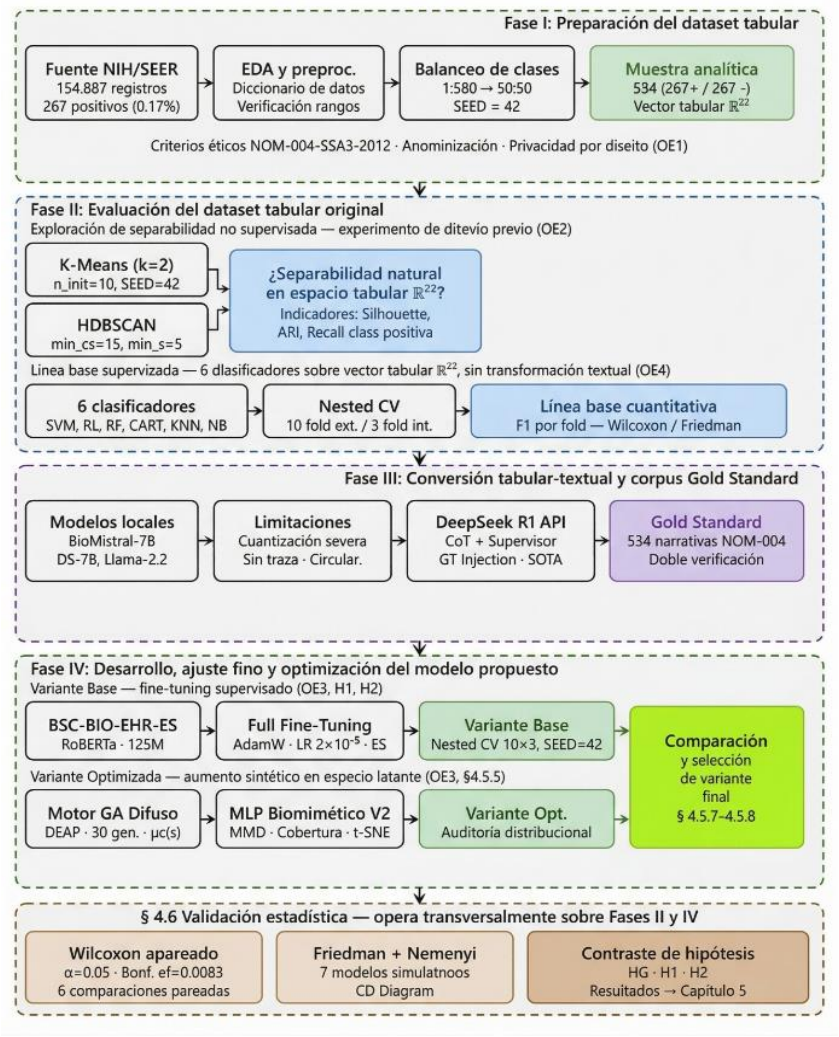

Figura 4.1: Flujo metodológico general: (I) preparación del dataset tabular → (II) evaluación con datos tabulares (separabilidad + baselines) → (III) conversión tabular-textual y Gold Standard → (IV) desarrollo, ajuste fino y optimización del modelo propuesto. La validación estadística (4.6) opera transversalmente sobre las Fases II y IV.

El orden de las fases responde a una lógica de justificación acumulativa. *La Fase I* garantiza que el corpus de trabajo es válido, balanceado y éticamente resguardado. *La Fase II* establece empíricamente los límites de lo que es alcanzable con la representación tabular original, tanto en modo no supervisado como con clasificadores supervisados; estos resultados constituyen la línea base contra la cual se medirá la aportación del enfoque propuesto. *La Fase III* produce el corpus en lenguaje natural que es condición arquitectónica para activar la capacidad del modelo *Transformer*. *La Fase IV* desarrolla, ajusta y optimiza el modelo sobre ese corpus, contrastando una variante base y una variante aumentada para seleccionar la configuración final.

4.2 Fase I. Preparación del *Dataset* clínico tabular

4.2.1 Fuente del *Dataset* y estructura tabular original

Con el propósito de disponer de un corpus clínico verificado y reproducible para la investigación en cumplimiento con el *OE1*, el conjunto de datos base fue obtenido del Instituto Nacional de Salud de los Estados Unidos (*NIH*) a través del programa *SEER* de acceso a datos de investigación oncológica (*National Institutes*

of Health, "Surveillance, Epidemiology, and End Results (SEER) Program", 2023). Los registros corresponden a una cohorte de individuos con diagnóstico confirmado de cáncer de tiroides y controles sanos, recopilados bajo protocolos aprobados por comités institucionales de ética. La fuente fue seleccionada por su disponibilidad abierta bajo licencia de investigación y por la presencia de diagnósticos verificados que confieren validez al estándar de oro de las etiquetas de clase.

El archivo original contiene 154,887 registros y 23 columnas: 22 variables predictoras y una variable objetivo binaria. La heterogeneidad de escalas combinando variables continuas, ordinales y nominales es un rasgo que se aborda explícitamente en el preprocesamiento de la *Sección 4.2.6*, y que motiva además la evaluación de representaciones alternativas documentada en las fases siguientes.

4.2.2 Variables clínicas, codificación y naturaleza de los datos

La tabla 4.1 describe las 22 variables predictoras con su tipo de escala, codificación y relevancia clínica para el diagnóstico de cáncer de tiroides.

Tabla 4.1. Descripción de las 22 variables clínicas del dataset NIH y su codificación.

#	Variable	Escala	Categorías / Rango	Relevancia
1	Sexo biológico	Nominal dicotómica	0=Femenino, 1=Masculino	Factor de riesgo
2	Edad al diagnóstico (años)	Razón continua	18–85 años	Factor de riesgo
3	Raza/etnicidad	Nominal *polic*. (1–5)	1=Blanco, 2=Negro, 3=Hispano., 4=Asiático, 5=Otro	Factor demográfico
4	Nivel educativo	Ordinal (1–6)	1=Sin estudios … 6=Posgrado	Determinante social
5	Ocupación	Nominal (1–5)	1=Ama de casa … 5=Profesional	Determinante social
6–7	Peso (libras) y talla (pulgadas)	Razón continua	≥80 lbs / ≥48 pulgadas	Medida antropométrica
8–10	*IMC* actual, a los 20 y a los 50 años	Ordinal (1–4)	1=Bajo, 2=Normal, 3=Sobrepeso, 4=Obesidad	Historial metabólico
11	Hábito tabáquico	Nominal (0–2)	0=Nunca, 1=Ex-fumador, 2=Activo	Factor de riesgo
12–13	Antecedente familiar: cáncer tiroides / otro cáncer	Nominal dicotómica	0=No, 1=Sí	Historial genético
14–22	Variables de comorbilidad y exposición ambiental	Nominal / Ordinal	Diversas codificaciones	Contexto clínico
23	Diagnóstico de cáncer de tiroides --- VARIABLE OBJETIVO	Nominal dicotómica	0=Negativo, 1=Positivo	Gold Standard

La coexistencia de variables continuas (peso, talla, edad), ordinales (*IMC* categórico, nivel educativo) y nominales sin orden natural (sexo, raza, ocupación) impide la aplicación directa de métricas de distancia euclidiana como criterio diagnóstico, lo que constituye uno de los hallazgos motivadores del enfoque semántico desarrollado en la *Fase IV*.

4.2.3 Criterios de inclusión, exclusión y depuración

Los criterios que rigen la composición de la muestra analítica son los siguientes:

Criterios de inclusión:

- Registro con las 22 variables predictoras completas, sin valores ausentes en los campos de clasificación críticos (sexo, edad, *IMC*, tabaquismo, antecedentes oncológicos).
- Para la clase positiva: presencia de la etiqueta diagnóstica confirmada (variable 23 = 1) en el registro *NIH* original.
- Registro que, tras el proceso de conversión textual (*Fase III*), genere una narrativa con puntuación de calidad ≥ 5/10 en la rúbrica del Agente Supervisor y sin discrepancias críticas entre el *Ground Truth* inyectado y el contenido generado.

Criterios de exclusión:

- Registro cuya narrativa generada presente alucinaciones extrínsecas mayores invención de diagnósticos, procedimientos o datos demográficos sin respaldo en el *CSV*, independientemente de su puntuación global.
- Registro con discrepancia irreconciliable en campos de alta relevancia clínica (sexo, edad, diagnóstico) que no pueda corregirse mediante regeneración.
- Registros duplicados identificados durante la auditoría del pipeline.

4.2.4 Balanceo de clases y construcción de la muestra analítica

El *Dataset* original de 154,887 registros presenta una distribución de clases fuertemente asimétrica: únicamente 267 instancias (0.17%) pertenecen a la clase positiva, lo que configura una proporción de 1:580 por cada caso de cáncer de tiroides existen 580 controles sanos. Esta condición de Desbalance Extremo (Chawla et al., 2002) invalida la exactitud (*Accuracy*) como métrica de evaluación: un clasificador que prediga sistemáticamente la clase negativa obtendría más del 99% de exactitud aparente sin haber aprendido ningún patrón clínico. La estrategia de depuración adoptada es el submuestreo balanceado: (i) extracción exhaustiva de los 267 casos positivos disponibles; (ii) submuestreo aleatorio estratificado de 267 casos negativos sin reemplazo (*SEED=42*). El resultado es la muestra analítica de 534 registros con distribución 50:50 que constituirá la base de todas las fases experimentales posteriores. Dado que los 267 casos positivos representan la totalidad de instancias disponibles en la fuente, la extensión del corpus de la clase minoritaria no puede ampliarse mediante otras técnicas sin recurrir a generación sintética, lo que establece empíricamente el techo del recurso de datos real disponible (Chawla et al., 2002). La figura 4.2 ilustra la transformación de la distribución de clases.

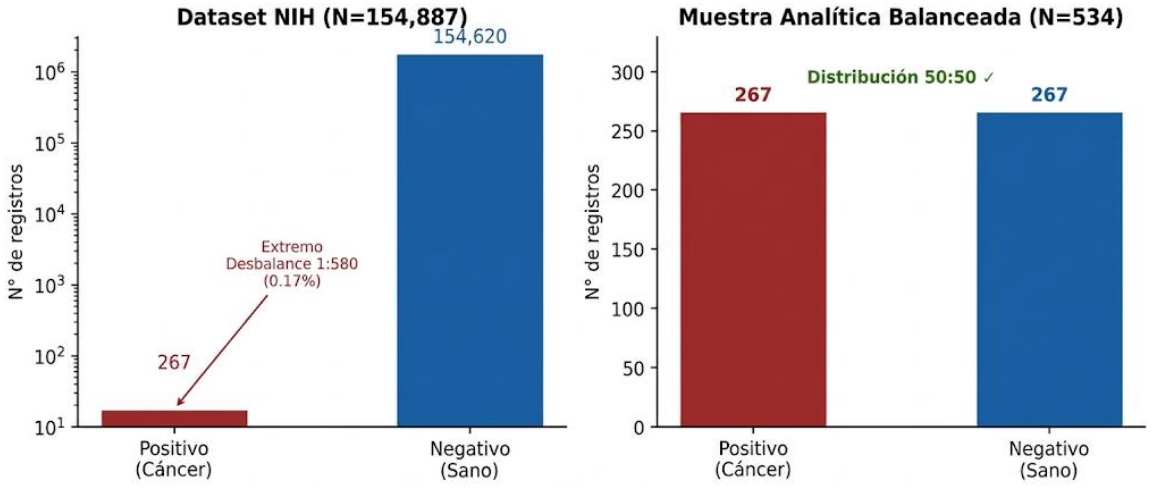

Figura 4.2: Distribución de clases: dataset original NIH (154,887 registros, proporción 1:580, escala logarítmica) versus muestra analítica balanceada (534 registros, proporción 1:1).

4.2.5 Anonimización, consideraciones éticas y resguardo de la información

En cumplimiento con el *OE1*, el presente estudio trabaja exclusivamente con datos de investigación de acceso abierto sin contacto con pacientes ni con expedientes clínicos primarios. Los mecanismos de resguardo adoptados son:

Anonimización operativa: las variables demográficas se expresan mediante categorías generales sin valores individuales exactos, en consistencia con las disposiciones de la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012) sobre información del expediente clínico sin exposición de datos de identidad.

Privacidad por diseño en el aumento sintético: la experimentación de aumento de datos (*sección 4.5.6*) opera exclusivamente sobre vectores matemáticos de 768 dimensiones, sin acceso ni procesamiento de texto clínico real durante la fase generativa.

Seguridad de la infraestructura: el *VPS* utilizado para el procesamiento fue sometido a un protocolo de endurecimiento que incluyó autenticación por llave asimétrica, migración del servicio *SSH* al puerto 5022, *firewall UFW* de mínimo privilegio, e intercambio de llaves *Curve25519* con cifrado *ChaCha20-Poly1305*. Este protocolo fue activado como respuesta a un incidente de seguridad detectado durante la fase exploratoria compromiso por exposición de puertos *SMB/Samba* ante la naturaleza médica del corpus.

Uso declarado de inteligencia artificial: de acuerdo con las directrices de *Ellaway* y *Tolsgaard* (R. Ellaway & Tolsgaard, 2023), se declara explícitamente que *DeepSeek R1*, accedido vía *API*, fue utilizado como herramienta de conversión textual, no como agente de decisiones metodológicas. El protocolo cumple los principios de Beneficencia, No Maleficencia, Autonomía y Justicia de la Declaración de *Helsinki* adaptados al contexto de investigación computacional con datos secundarios oncológicos.

4.2.6 Preprocesamiento del *Dataset* tabular para análisis y clasificación

El preprocesamiento del *Dataset* tabular se aplica en dos contextos distintos con objetivos diferentes, lo que determina procedimientos específicos para cada caso.

Para la *Fase II* (evaluación con datos tabulares): las 22 variables predictoras se codifican numéricamente en su forma original, sin transformación textual. Las variables nominales se convierten mediante codificación ordinal preservando el código numérico original del *Dataset*; las variables continuas se normalizan mediante estandarización (media cero, desviación estándar uno) para garantizar la comparabilidad de escalas en los algoritmos sensibles a magnitudes (*SVM, KNN*). Este preprocesamiento produce un vector de características $\chi \in \mathbb{R}^{22}$ por muestra, que constituye la representación de entrada directa para los seis algoritmos de la línea base y para los experimentos no supervisados.

Para la *Fase III* (conversión a lenguaje natural): el *Dataset* tabular es la fuente de los valores que se decodifican mediante el Diccionario de Datos para generar las narrativas clínicas *NOM-004*. En este contexto, el preprocesamiento incluye la validación de rangos numéricos, la verificación de la consistencia de los códigos categóricos y la generación del Diccionario de Datos documento de mapeo de cada columna *CSV* a su significado semántico descrito en detalle en la sección 4.4.

4.3 Fase II. Evaluación del *Dataset* tabular original

En cumplimiento con el *OE2*, la Fase II evalúa empíricamente las características del *Dataset* tabular original como representación de entrada para la clasificación. Esta evaluación se desarrolla en dos etapas secuenciales con lógica propia: primero se explora si los datos presentan separabilidad geométrica natural en ausencia de supervisión, y posteriormente se establece la línea base supervisada que servirá como referencia comparativa para el modelo propuesto.

4.3.1 Exploración preliminar de separabilidad natural en el espacio tabular

Antes de comprometer recursos en el ajuste fino de un modelo de lenguaje, se planteó la siguiente pregunta de diseño: ¿es posible que las 22 variables clínicas del *Dataset* tabular contengan estructuras de agrupamiento natural que permitan distinguir casos positivos de negativos sin necesidad de etiquetas ni de una representación textual? De existir tal separabilidad, el enfoque basado en lenguaje natural no sería estrictamente necesario, y podría sustituirse por métodos de agrupamiento directos sobre los datos tabulares.

Para responder esta pregunta se aplicaron dos algoritmos de agrupamiento no supervisado sobre la muestra analítica balanceada en su representación tabular directa. Estos experimentos tienen carácter exploratorio: no forman parte del flujo principal de entrenamiento del modelo, sino que constituyen evidencia para una decisión de diseño. Sus resultados se presentan en el Capítulo 5.

4.3.1.1 Evaluación exploratoria con *K-Means*

Se aplicó el algoritmo *K-Means* con *k=2* centroides forzados, correspondientes a las dos clases diagnósticas (sano y enfermo), con *n_init=10* y *random_state=42* para reproducibilidad. La configuración de dos centroides es la más favorable al escenario de separabilidad que se desea probar: si bajo esta condición el agrupamiento no logra alinear sus *clústers* con las etiquetas reales, la separabilidad geométrica esférica queda descartada. Los indicadores calculados *Silhouette Score*, *Adjusted Rand Index* (*ARI*) y tasa de recuperación de la clase positiva se reportan en el Capítulo 5.

4.3.1.2 Evaluación exploratoria con *HDBSCAN*

Se aplicó el algoritmo *HDBSCAN* con *min_cluster_size=15* y *min_samples=5* para evaluar si existe separabilidad basada en densidad, independientemente de la geometría esférica de *K-Means*. *HDBSCAN* identifica *clústers* de forma arbitraria sin requerir la especificación a priori del número de grupos, lo que lo hace idóneo para detectar estructuras de agrupamiento que podrían no seguir distribuciones regulares. La tasa de recuperación de la clase positiva, el *ARI* y la proporción de puntos clasificados como ruido son los indicadores principales de esta evaluación, documentados en el Capítulo 5.

4.3.1.3 Alcance metodológico de los experimentos no supervisados

Ambos experimentos cumplen una función metodológica específica y acotada: determinar si el problema admite un enfoque de descubrimiento de estructura latente en el espacio tabular como alternativa o complemento al enfoque supervisado. Su inclusión en el diseño responde a la premisa de no asumir a priori que el espacio tabular carece de información estructural aprovechable; esa premisa debe ser verificada empíricamente, no supuesta. Los resultados de esta etapa exploratoria informan la interpretación de los algoritmos supervisados que se describen a continuación.

4.3.2 Línea base comparativa con algoritmos de clasificación tradicionales

En el marco del *OE4*, que establece la necesidad de comparar el modelo propuesto con métodos de referencia, se diseñó un protocolo de línea base supervisada con seis algoritmos de clasificación clásicos que operan directamente sobre el *Dataset* tabular. Esta evaluación produce los valores de *F1* por pliegue que servirán como insumo para las pruebas estadísticas de la sección 4.6 y habilita la propuesta de la Hipótesis General.

4.3.2.1 Justificación de los modelos *Baseline*

Los clasificadores clásicos seleccionados cumplen cuatro funciones metodológicas: (i) proporcionar una línea base cuantitativa objetiva contra la cual contrastar el modelo de lenguaje ajustado; (ii) establecer el máximo rendimiento alcanzable desde la representación tabular directa sin transformación semántica; (iii) generar los vectores de métricas por pliegue necesarios para las pruebas de *Wilcoxon* y *Friedman*; y (iv) verificar empíricamente si el enfoque propuesto representa una aportación diferencial sobre el estado clásico del arte para este dominio. La evaluación de la línea base antecede cronológicamente al diseño del experimento de fine-tuning: el nivel de rendimiento observable con enfoques clásicos es un dato de diseño que contextualiza la dificultad del problema.

4.3.2.2 Algoritmos seleccionados

Se seleccionaron seis algoritmos que representan las principales familias teóricas del aprendizaje automático supervisado (Pedregosa et al., 2012):

• Máquina de Vectores de Soporte (*SVM, Kernel RBF*): clasificador de margen máximo con proyección implícita a espacios de alta dimensión. Referencia de la capacidad de separación no lineal basada en márgenes.

• *Regresión Logística* con regularización *L2*: clasificador lineal interpretable. Referencia de la capacidad máxima de separación lineal del espacio tabular.

• *Random Forest* (100 estimadores): conjunto de árboles con muestreo aleatorio de características. Referencia de la capacidad de capturar interacciones no lineales entre variables clínicas.

• Árbol de Decisión (*CART*): partición recursiva del espacio de características. Incluido por su interpretabilidad completa y como límite inferior del rendimiento no paramétrico simple.

• *K-Vecinos* Más Cercanos (*KNN, k=5*, distancia euclidiana): clasificador por proximidad local. Permite evaluar directamente la hipótesis de separabilidad por cercanía entre vectores de pacientes.

• *Naive Bayes Gaussiano*: clasificador probabilístico bajo independencia condicional. Límite inferior teórico bajo el supuesto más restrictivo de independencia entre variables.

4.3.2.3 Representación de entrada y preprocesamiento tabular

Todos los algoritmos reciben como entrada el vector tabular de 22 características $\chi \in \mathbb{R}^{22}$ descrito en la sección 4.2.6, que corresponde al *Dataset* original en su estado tabular preprocesado: variables continúas estandarizadas y variables categóricas codificadas numéricamente. No se aplica ninguna transformación textual ni representación de bolsa de palabras. Esta configuración es la comparación directa y honesta entre el dato tabular como fuente y la narrativa clínica como fuente: los clasificadores clásicos reciben la información en su forma estructurada original, mientras que el modelo de lenguaje la recibe convertida a texto. Cualquier diferencia de rendimiento observada entre ambos enfoques es atribuible a la representación, no a ventajas de datos.

4.3.2.4 Protocolo de entrenamiento y validación

Los seis algoritmos se entrenan y evalúan bajo el mismo esquema de Validación Cruzada Anidada (10 pliegues externos / 3 pliegues internos, *SEED=42*) definido para el modelo de lenguaje en la sección 4.5.4, garantizando la comparabilidad directa de los resultados. La implementación se realiza con *scikit-learn* (Pedregosa et al., 2012) mediante la interfaz *Pipeline* que encadena el preprocesamiento tabular y el clasificador. Los *hiperparámetros* de cada algoritmo se fijan en sus valores por defecto de *scikit-learn* para garantizar la objetividad de la comparación y evitar que su optimización favorezca artificialmente a los algoritmos clásicos.

4.3.2.5 Métricas de comparación

Las métricas calculadas para cada algoritmo son: *F1-Score* (métrica principal de comparación), *ROC-AUC*, *Accuracy*, *Precisión* y *Recall*, calculadas sobre los 10 pliegues externos del *Nested CV*. Los valores de *F1* por pliegue de cada modelo constituyen el vector de observaciones apareadas que se usará en las pruebas de *Wilcoxon* y *Friedman* de la sección 4.6. Los resultados de esta evaluación se documentan en el Capítulo 5.

4.4 Fase III. Conversión tabular-textual y construcción del corpus clínico *Gold Standard*

4.4.1 Fundamentación de la transformación tabular a lenguaje natural

Los modelos *Transformer* de clasificación de secuencias como *BSC-BIO-EHR-ES* procesan secuencias de tokens lingüísticos, no vectores numéricos. Su mecanismo de autoatención (Vaswani et al., 2017) y su *tokenizador* especializado operan sobre texto; el conocimiento médico adquirido durante el preentrenamiento solo se activa cuando la entrada adopta el formato lingüístico sobre el que el modelo fue entrenado. La

conversión tabular-textual es, por tanto, una condición arquitectónica necesaria para poder hacer uso de la capacidad de comprensión clínica del modelo, no una opción de diseño sustituible.

La calidad de esta conversión tiene consecuencias directas sobre la integridad experimental. Una narrativa con alucinaciones clínicas datos fabricados sin respaldo en el *CSV* introduciría ruido semántico que contaminaría el corpus de entrenamiento con asociaciones espurias entre patrones lingüísticos y etiquetas diagnósticas. Por ello, el pipeline de conversión fue diseñado con verificación de doble instancia, cuyo protocolo se describe en la sección 4.4.6.

El formato adoptado es la historia clínica electrónica según la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012), en cumplimiento con el *OE1*. Esta elección garantiza la coherencia lingüística con el dominio objetivo (sistema de salud en español) y maximiza la activación del conocimiento preentrenado de un modelo entrenado con *EHR* en español. La figura 4.3 ilustra el flujo completo de conversión.

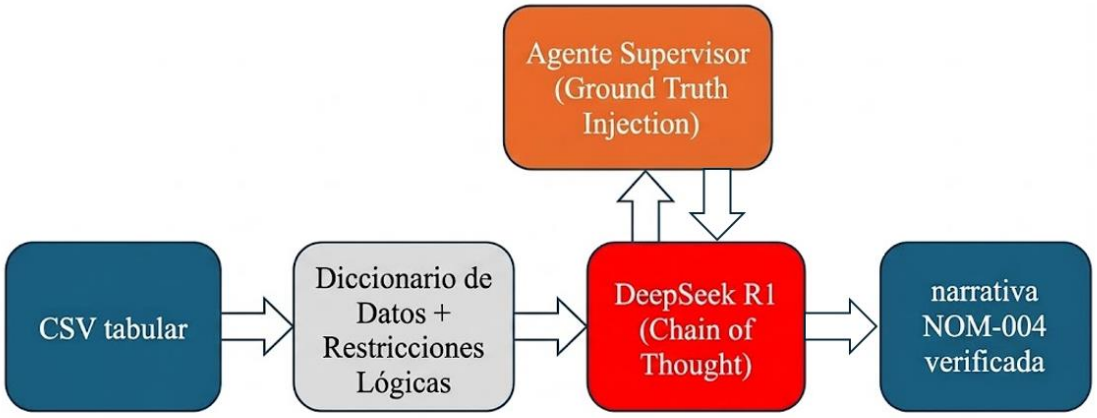

Figura 4.3: Flujo de transformación tabular-textual: CSV tabular → Diccionario de Datos + Restricciones Lógicas → DeepSeek R1 (Chain of Thought) → Agente Supervisor (Ground Truth Injection) → narrativa NOM-004

4.4.2 Exploración preliminar con modelos locales

La estrategia inicial exploró el despliegue de modelos de lenguaje en el *VPS* local (12 *vCPU*, 48 *GB RAM*) para evitar los costos de la inferencia por *API*. Se evaluaron tres modelos en formato *GGUF* cuantizado mediante llama.cpp. Para cada uno se ejecutaron entre 8 y 13 pruebas sobre el mismo registro de referencia, calificadas mediante una rúbrica holística de 0–10/10.

4.4.2.1 *BioMistral-7B*

BioMistral-7B (*7B* parámetros, *Q4_K_M*, preentrenado con *PubMed*) fue evaluado con *n_ctx=8,192*, *Top_p=0.95*, *Top_k=40* y temperatura de 0.7 a 0.0. Los resultados evidenciaron tres categorías de fallo: alucinaciones extrínsecas (invención de protocolos oncológicos *absent* del *CSV*), errores fácticos en campos fundamentales (unidades y datos demográficos incorrectos) y fallo de instrucción (regurgitación literal del *prompt*). El diagnóstico causal indica que el preentrenamiento narrativo con publicaciones médicas predispone al modelo a completar historias clínicas creativamente, no a decodificar valores tabulares. Tasa de éxito: 14%. Calificación promedio: 1.1/10.

4.4.2.2 *DeepSeek-7B*

DeepSeek-7B (*7B* parámetros, propósito general, *temperature=0.0*) fue evaluado bajo la premisa de que la ausencia de sesgo médico favorecería la decodificación fiel de datos. Los resultados mostraron cuatro incapacidades: analfabetismo estructural ante *CSV* (razonamiento conceptual en lugar de *parseo*), interpretación

literal de códigos categóricos (el código 5 de educación procesado como edad de 5 años), fallo catastrófico del aprendizaje *few-shot* (mezcla de datos del ejemplo con el registro real) y desobediencia algorítmica (cálculo del *IMC* a pesar de la prohibición explícita). Tasa de éxito: 0%. Calificación promedio: 0.3/10.

4.4.2.3 *Llama-3.2-3B-Instruct*

Llama-3.2-3B-Instruct (*3B* parámetros, variante *Instruct*, *Q5_K_M*) obtuvo la mejor tasa relativa (62.5% de éxito, calificación máxima 6/10), pero presentó limitaciones no resolubles mediante ajuste del *prompt*: el código de *IMC* Normal fue interpretado sistemáticamente como Obesidad (100% de fallos), la categoría Ama de Casa no fue identificada correctamente en ninguna ejecución, y el módulo de validación posterior a la generación (primer Filtro) falló en el 25% de las ejecuciones. Este último fallo evidenció que emplear el mismo modelo para generar y verificar introduce dependencia circular: los errores sistemáticos del generador se replican en la fase de validación.

4.4.3 Limitaciones detectadas en la inferencia local

La evaluación sistemática permitió identificar cuatro limitaciones estructurales que motivaron el rediseño de la estrategia:

- Degradación cognitiva por cuantización severa (*Q4*–*Q5*): la compresión de los pesos deteriora la capacidad de razonamiento estructurado, especialmente en tareas que requieren interpretar instrucciones complejas y relaciones numéricas entre campos.
- Conflicto entre preentrenamiento narrativo y tarea decodificadora: los modelos aprendieron a generar texto fluido, no a actuar como traductores fieles de estructuras de datos. Este conflicto no se resuelve mediante la optimización del *prompt* porque está integrado en los pesos del modelo.
- Imposibilidad de verificación asimétrica: utilizar el mismo modelo para generar y verificar crea dependencia circular, como demostró el fallo del primer Filtro de *Llama*.
- Ausencia de traza de razonamiento: sin la exposición del proceso lógico intermedio, la corrección dirigida de errores específicos es imposible.

4.4.4 Selección de *DeepSeek R1* vía *API* como estrategia definitiva

Las limitaciones identificadas condujeron a la adopción de la inferencia vía *API* con *DeepSeek R1* en modalidad *SOTA* (D. Guo et al., 2025b). Esta selección resuelve las cuatro limitaciones estructurales: la capacidad de razonamiento del modelo *SOTA* no está degradada por cuantización; su preentrenamiento de escala superior le permite procesar datos estructurados; la arquitectura de doble instancia del *pipeline* (generador separado del verificador) elimina la dependencia circular; y la Traza de Razonamiento expuesta permite la auditoría del proceso lógico intermedio. Es metodológicamente relevante distinguir este modelo del *DeepSeek-7B* cuantizado evaluado en 4.4.2.2: la denominación compartida no implica equivalencia; el modelo *SOTA* accedido vía *API* tiene capacidades cualitativamente superiores. Frente a alternativas de similar potencia (*GPT-4o*, *Claude 3.5 Sonnet*), *DeepSeek R1* (Guo et al., 2025b) presenta dos ventajas determinantes: costo operativo bajo que permite múltiples iteraciones de generación y verificación, y exposición de la Traza de Razonamiento (*Reasoning Trace*), propiedad ausente en los modelos competidores. La tabla 4.2 presenta la comparativa.

Tabla 4.2. Comparativa de motores de inferencia vía API para la conversión clínica. ★ Modelo seleccionado. La *Reasoning Trace* es la propiedad crítica que habilita la arquitectura de verificación asimétrica del pipeline (Guo et al., 2025b).

Modelo API	Costo relativo	Traza de razonamiento	Consistencia tabular	Política de privacidad
DeepSeek R1 ★	Muy bajo	Alta --- Reasoning Trace expuesta	Alta	Pública y verificable
GPT-4o	Muy alto	No disponible	Alta	Restrictiva
Gemini 1.5 Pro	Moderado	No disponible	Variable	Restrictiva
Claude 3.5 Sonnet	Alto	No disponible	Alta	Verificable

4.4.5 Protocolo de generación textual en modo *SOTA*

El protocolo definitivo se fundamenta en tres pilares que resuelven las limitaciones de la fase exploratoria:

Pilar 1 --- Optimización de *Prompts*:

Los *System Prompts* incorporan tres módulos funcionales: (a) Diccionario de Datos con mapeo explícito de cada columna *CSV* a su significado semántico y leyendas de decodificación completas para variables categóricas; (b) Restricciones lógicas condicionales con reglas para valores nulos, campos sensibles y prohibiciones de cálculo derivado; y (c) Directivas de estilo oncológico *NOM-004* con registro formal del lenguaje médico y penalización explícita de alucinaciones.

Pilar 2 --- *Chain of Thought (CoT)*:

El mecanismo *CoT* instruye al modelo a descomponer la tarea en pasos lógicos verificables antes de generar la narrativa. La propiedad crítica de *DeepSeek R1* es la traza de razonamiento, que convierte la inferencia de opaca en auditable: el agente supervisor puede revisar cada paso y corregir desviaciones antes de que se materialicen en errores factuales.

Pilar 3 --- Agente Supervisor con Ground Truth Injection:

Flujo de doble instancia asimétrica: (a) el modelo principal genera la narrativa usando *CoT*; (b) el Agente Supervisor recibe la narrativa y los valores crudos del *CSV* inyectados mediante etiquetas *XML (<edad>74</edad>, <sexo>Femenino</sexo>)*. La regla de prioridad ontológica 'los marcadores gobiernan sobre la narrativa' determina que toda discrepancia activa la reescritura selectiva del fragmento afectado. La asimetría de roles el Supervisor solo valida, no genera elimina la dependencia circular del primer filtro.

4.4.6 Supervisión, validación y control de fidelidad semántica

El control de calidad del corpus opera en dos instancias secuenciales. La primera es el filtrado automático por el *Agente Supervisor*, que verifica campo por campo la correspondencia entre el *Ground Truth* inyectado y la narrativa generada para los campos de mayor relevancia clínica. La segunda es la auditoría manual de una muestra representativa, donde el investigador verifica la congruencia narrativa global, la ausencia de contradicciones lógicas sutiles y la fidelidad al registro formal de la *NOM-004*.

Los tipos de error clasificados según severidad son: *Tipo A* alucinaciones extrínsecas mayores (invención de datos sin respaldo en el *CSV*), motivo de descarte inmediato; *Tipo B* errores fácticos menores en campos de baja relevancia, corregibles por el *Agente Supervisor*; y *Tipo C* imprecisiones narrativas de estilo que no alteran el significado clínico, admisibles.

4.4.7 Consolidación del corpus clínico en lenguaje natural

El corpus *Gold Standard* resultante comprende 534 registros (267 positivos, 267 negativos), cada uno representado como un par (*texto_narrativo, etiqueta_diagnóstica*). Cada narrativa tiene entre 150 y 300 palabras y cumple el formato de historia clínica *NOM-004-SSA3-2012*. El corpus se almacenó como archivo *CSV* con columnas *text* y *label*, constituyendo el único insumo de entrenamiento para la *Fase IV*.

4.5 *Fase IV*. Desarrollo, ajuste y optimización del modelo propuesto

La *Fase IV*, que da respuesta directa al *OE3*, comprende la selección del modelo base, el diseño y ejecución del ajuste fino supervisado, y la experimentación de aumento de datos en espacio latente como vía de optimización. El resultado de esta fase es una variante base y una variante optimizada del modelo, cuya comparación y selección se documentan en las secciones 4.5.7 y 4.5.8.

4.5.1 Criterios de selección del modelo base

La selección del modelo base se realizó mediante la operacionalización de tres dimensiones de requerimiento:

Dimensión clínica:

Los principios de Beneficencia y No Maleficencia se traducen en dos requerimientos métricos sobre el modelo candidato: alta capacidad de detección de entidades oncológicas (Requerimiento A, operacionalizado por el F1 en el benchmark CANTEMIST) y alta comprensión de informes clínicos reales en español (Requerimiento B, operacionalizado por el F1 en el benchmark ICTUSnet). El F1-Score del modelo ajustado como media armónica de Precisión y Recall servirá como indicador maestro de ambos requerimientos simultáneamente.

Dimensión lingüística:

El modelo debe haber sido preentrenado en español biomédico. Los modelos entrenados en inglés o en español general presentan fragmentación inadecuada de términos clínicos en el tokenizador: un término como 'Tiroides' se segmenta en cuatro sub-tokens sin significado médico en un tokenizador de propósito general, frente a un token completo en uno especializado. Esta fragmentación diluye las representaciones vectoriales de términos oncológicos clave (Carrino et al., 2021).

Dimensión computacional:

Criterios de seleccionabilidad: código abierto con pesos en repositorio público (reproducibilidad, OE1); arquitectura encoder-only para clasificación discriminativa; tamaño compatible con el hardware disponible. Criterio de exclusión: política de acceso restringido que impida la reproducción del experimento.

4.5.2 Justificación de la selección de BSC-BIO-EHR-ES

El modelo BSC-BIO-EHR-ES del Plan de Tecnologías del Lenguaje del Gobierno de España (Carrino et al., 2021) fue seleccionado tras la evaluación comparativa documentada en la tabla 4.3. MédicoBERT (UAEMex) fue descartado por política de acceso restringido incompatible con los criterios de reproducibilidad del OE1.

Tabla 4.3. Evaluación comparativa de modelos de lenguaje biomédico en español. Fuente: Valores de benchmarks públicos de Carrino et al. (Carrino et al., 2021); no son resultados del presente estudio. ★ Modelo seleccionado.

Modelo	Arquitectura base	CANTEMIST (F1)	ICTUSnet (F1)	Acceso
bsc-bio-ehr-es ★	RoBERTa Bio+EHR Español	0.834	0.8756	Abierto
XLM-R-Galén	XLM-R Multilingüe	0.8078	0.8716	Abierto
BETO-Galén	BERT Español General	0.8153	0.8498	Abierto
mBERT	BERT Multilingüe	0.8116	0.8631	Abierto
MédicoBERT(UAEMex)	BERT Español Médico MX	N/D	N/D	Restringido

BSC-BIO-EHR-ES es la única opción disponible que satisface simultáneamente el Requerimiento A y el Requerimiento B con el mayor nivel de rendimiento preentrenado. Su preentrenamiento con 95 millones de tokens de Registros Electrónicos de Salud (EHR) reales en español le aporta la capacidad semántica de distinguir entre antecedente familiar de riesgo y diagnóstico activo distinción crítica para minimizar falsos positivos en el contexto oncológico tiroideo, que ningún modelo de español general puede replicar. La arquitectura RoBERTa (Liu et al., 2019), al prescindir de la predicción de la oración siguiente y emplear máscaras dinámicas, produce representaciones semánticas más robustas para clasificación de secuencias largas.

4.5.3 Diseño experimental del fine-tuning supervisado

4.5.3.1 Variables e indicadores

Variable independiente principal: estrategia de *Full Fine-Tuning* con la configuración de la tabla 4.4, sobre el corpus *Gold Standard* de 534 registros en lenguaje natural.

Variables dependientes primarias: *F1-Score* (indicador maestro de *HG*), desviación estándar del *F1* entre pliegues (indicador de *H1*), curvas *Train/Val Loss* por época (indicador de *H2*), *ROC-AUC*, *Precisión*, *Recall* y *MCC*.

Variables de calidad probabilística: *ECE* (*Expected Calibration Error*) y *Brier Score* como indicadores de honestidad probabilística (C. Guo et al., 2017), relevantes para la evaluación del sistema como herramienta de apoyo a la decisión clínica (*OE5*).

Variables de control: *Dataset* fijo (*N=534, 50:50, SEED=42*), arquitectura lineal sobre el *token* [*CLS*] (Devlin et al., 2019), *tokenizador AutoTokenizer* de *BSC-BIO-EHR-ES*, *max_length=128*, esquema *StratifiedKFold* (*n_outer=10*, *n_inner=3*, *random_state=42*).

4.5.3.2 Preparación de datos para entrenamiento y validación

Las narrativas del *Gold Standard* se procesan mediante el *AutoTokenizer* de *BSC-BIO-EHR-ES*, con *padding='max_length'*, *truncation=True*, *max_length=128 tokens* (Wolf et al., 2020). La longitud de 128 fue seleccionada por cubrir el rango de extensión de las narrativas *NOM-004* (extensión promedio 85–120 *tokens*) sin incurrir en los costos cuadráticos del mecanismo de atención para secuencias más largas (Vaswani et al., 2017). El *tokenizador* produce tres tensores por muestra: *input_ids*, *attention_mask* y *token_type_ids*. La división en pliegues estratificados garantiza la proporción 50:50 en cada subconjunto de entrenamiento y validación.

4.5.3.3 Configuración de *hiperparámetros*

Tabla 4.4. Configuración de hiperparámetros del protocolo de fine-tuning con justificación.

Hiperparámetro	**Valor**	**Justificación**
Optimizador	*AdamW*	Desacopla *Weight Decay* del gradiente, corrigiendo el error de regularización de *Adam* estándar para modelos de lenguaje (Loshchilov & Hutter, 2019).
Tasa de aprendizaje	2×10^{-5}	Rango estándar para *fine-tuning BERT/RoBERTa* (Devlin et al., 2019)(Howard & Ruder, 2018). Corresponde a *H2*. Valores $>5\times10^{-5}$ se asocian a inestabilidad en regímenes de pocos datos (Mosbach et al., 2021).
Batch Size	16	Recomendado para *fine-tuning* estable en *Datasets* pequeños (Mosbach et al., 2021). Equilibrio entre estabilidad del gradiente y restricciones de memoria.
Épocas máximas	6	Margen amplio para que el *Early Stopping* actúe. Convergencia típica documentada en 2–4 épocas (Devlin et al., 2019).
Early Stopping	Paciencia = 2	Regularizador efectivo en datos limitados (Caruana et al., 2000). Paciencia 2 balancea sensibilidad ante sobreajuste y riesgo de parada prematura.
Weight Decay	0.01	Regularización *L2* desacoplada estándar para modelos de lenguaje (Loshchilov & Hutter, 2019). Penaliza pesos grandes sin afectar parámetros de sesgo.
Longitud máxima	128 *tokens*	Cubre el percentil 95 de extensión de narrativas *NOM-004* sin costos cuadráticos del mecanismo de atención (Vaswani et al., 2017).
Métrica *Early Stopping*	*F1-Score*	Media armónica de *Precisión* y *Recall*; más informativa que *Accuracy* en *Datasets* balanceados como criterio de detección de sobreajuste.

4.5.3.4 Procedimiento de *fine-tuning*

El procedimiento se implementa mediante la *API Trainer* de *HuggingFace* (Wolf et al., 2020) con *SEED=42*. Para cada pliegue externo, el modelo se inicializa desde los pesos preentrenados limpios sin congelación de capas, garantizando la independencia entre evaluaciones. El ciclo por época comprende: (1) *forward pass* y cálculo de pérdida de entropía cruzada; (2) *backward pass* y actualización de parámetros con *AdamW*; (3) evaluación sobre el conjunto de validación; (4) registro del *F1* para el monitoreo de *Early Stopping*; (5) guardado del *checkpoint* con mejor *F1*. El proceso se interrumpe cuando el *F1* de validación no mejora durante dos épocas consecutivas. La figura 4.4 esquematiza el pipeline, mientras que en la figura 4.5 se visualizan las curvas de aprendizaje por *Fold*.

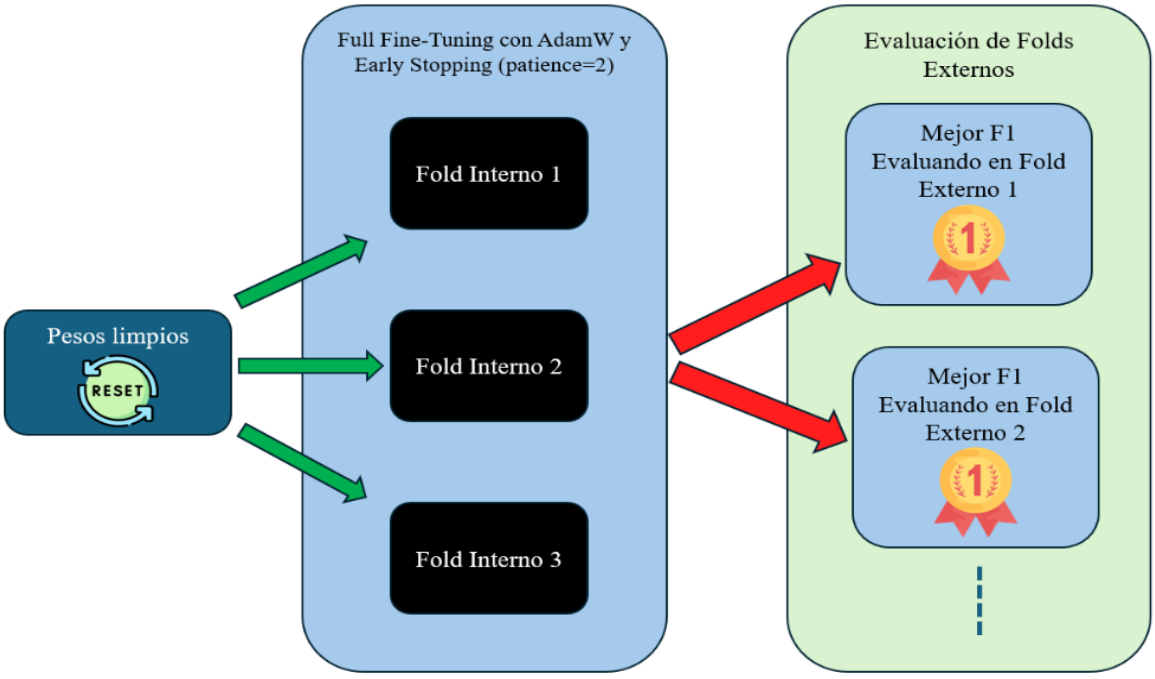

Figura 4.4: Pipeline de fine-tuning: inicialización desde pesos limpios por fold → Full Fine-Tuning con AdamW y Early Stopping (paciencia=2) → checkpoint con mejor F1 → evaluación en pliegue externo.

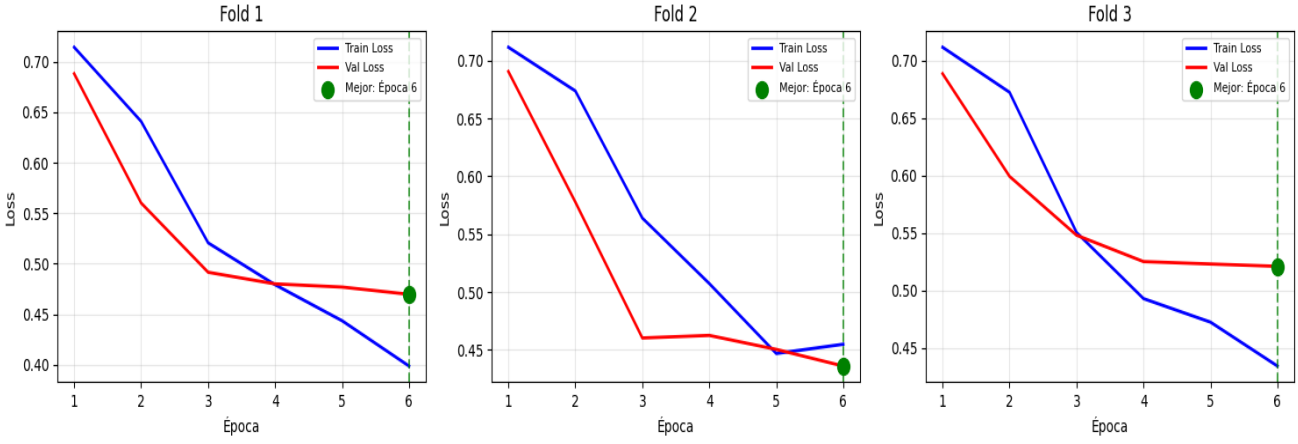

Figura 4.5: Esquema de curvas de aprendizaje por Fold: evolución de Train Loss y Val Loss por época. La dinámica relativa entre ambas curvas es el indicador de convergencia (H2) y de ausencia de memorización.

4.5.4 Esquema de *Nested Cross-Validation*

La validación sigue el protocolo de *Validación Cruzada* Anidada con 10 pliegues externos y 3 internos, conforme a las recomendaciones de *Forman* y *Scholz* (Forman & Scholz, 2010) para regímenes de datos limitados. El bucle externo se dedica exclusivamente a la estimación del rendimiento sobre datos no vistos; el bucle interno gestiona el entrenamiento y la selección del mejor *checkpoint*. Esta separación garantiza que el rendimiento reportado no está contaminado por el proceso de optimización (Demšar, 2006)(Forman & Scholz, 2010).

La discrepancia entre el rendimiento promedio del bucle interno y el del externo es el indicador técnico clave de generalización: una diferencia menor al 2–3% indicaría que el modelo produce estimaciones representativas de su comportamiento ante datos inéditos. La figura 4.6 ilustra el esquema de evaluación por pliegue externo.

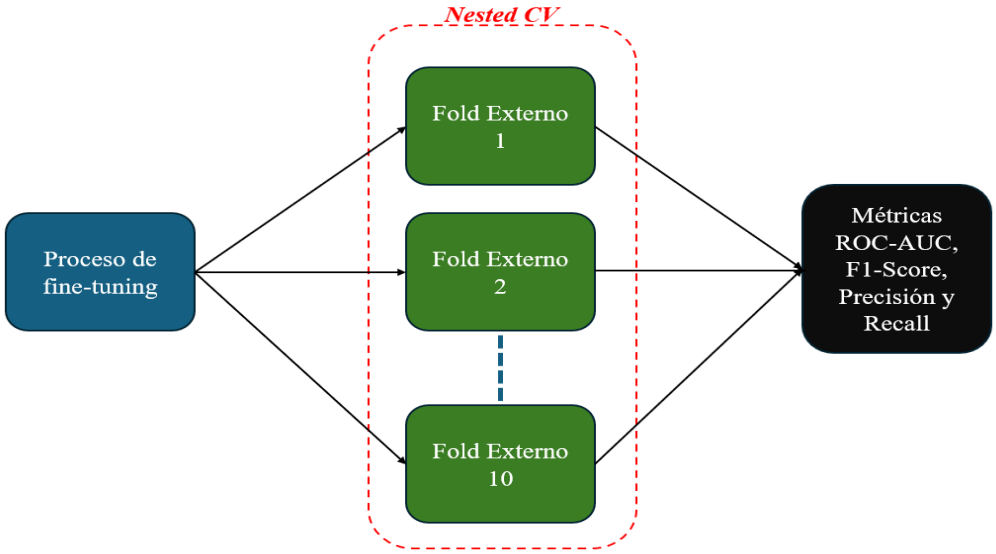

Figura 4.6: Esquema de evaluación por Fold externo bajo Nested CV (10 pliegues): distribución esperada de ROC-AUC, F1-Score, Precisión y Recall.

4.5.5 Métricas de desempeño, calibración e interpretabilidad

Las métricas cubren tres dimensiones del desempeño del sistema (*OE5*):

Rendimiento discriminativo: *F1-Score* (métrica primaria, balance clínico), *ROC-AUC* (potencia discriminativa independiente del umbral τ), *Precisión* (*VP/(VP+FP)*, indicador de No Maleficencia), *Recall* (*VP/(VP+FN)*, indicador de Beneficencia), y *MCC* (métrica robusta que considera los cuatro cuadrantes de la matriz de confusión).

Calidad probabilística: *Brier Score* (error cuadrático entre probabilidades predichas y etiquetas reales) y *ECE* (C. Guo et al., 2017) (diferencia promedio ponderada entre probabilidades predichas y frecuencias observadas). La calibración es crítica para un sistema de apoyo diagnóstico: las probabilidades emitidas serán interpretadas por médicos como grados de certeza, por lo que deben ser fidedignas (C. Guo et al., 2017). El análisis incluirá la curva de calibración con 10 *bins* bajo *Nested CV*.

Interpretabilidad (*OE5*): se aplicará *Feature Attribution* mediante modelo *surrogate* (*Random Forest* sobre representaciones internas del *Transformer*) para cuantificar la contribución relativa de los *embeddings* profundos frente a representaciones de frecuencia superficial. Adicionalmente se ejecutarán pruebas de estrés

lingüístico (paráfrasis, cambio de orden, negación local, sinónimos médicos) para caracterizar la robustez del modelo ante variaciones textuales semánticamente equivalentes.

4.5.6 Optimización bajo escasez de datos mediante aumento sintético en espacio latente

En el marco del *OE3*, y reconociendo que el corpus de 534 muestras establece un techo teórico a la capacidad de generalización del modelo, esta sección documenta la estrategia de aumento de datos en espacio latente diseñada para explorar si es posible mejorar el rendimiento del modelo base sin ampliar el corpus textual.

4.5.6.1 Justificación del aumento sintético

La obtención de datos clínicos adicionales en español conformes con *NOM-004* es inviable dentro de las restricciones del presente estudio (privacidad, normativa de acceso, disponibilidad institucional). El Aumento de *Datos* en *Espacio Latente* opera sobre las representaciones vectoriales del modelo preentrenado en lugar del texto superficial, con tres ventajas: integridad semántica (evita alucinaciones médicas en texto generado), eficiencia computacional (vectores numéricos se procesan órdenes de magnitud más rápido que reentrenamiento textual) y privacidad por diseño (las abstracciones vectoriales no exponen texto clínico real). Esta línea de experimentación produce la variante optimizada del modelo que será comparada con la variante base en la sección 4.5.7.

4.5.6.2 Arquitectura del sistema de aumento

La arquitectura consta de tres etapas modulares con roles no solapados. La figura 4.7 la ilustra.

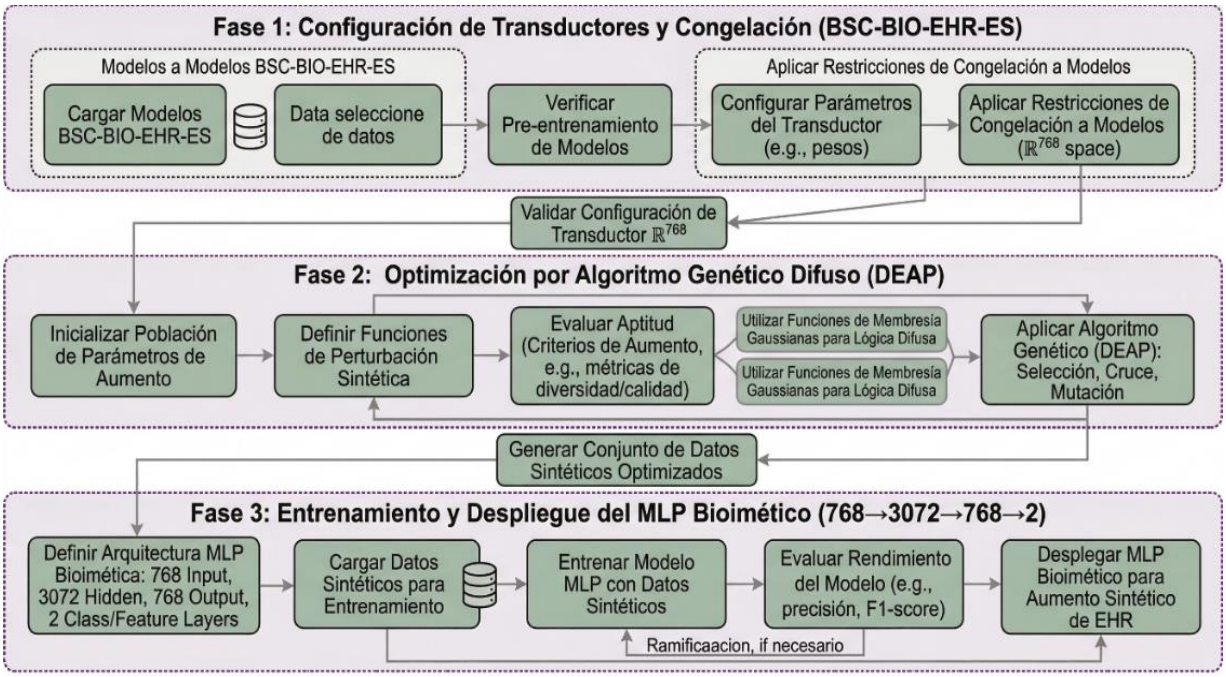

Figura 4.7: Arquitectura de tres etapas del sistema de aumento sintético: Etapa 1, BSC-BIO-EHR-ES congelado como transductor $\mathbb{R}^{768}$; Etapa 2, Algoritmo Genético Difuso (DEAP) con función de membresía gaussiana; Etapa 3, MLP Bioimético 768→3072→768→2.

Etapa 1 --- Extractor Semántico Estático: *BSC-BIO-EHR-ES* con pesos congelados en modo evaluación, actuando como transductor determinista que convierte cada narrativa en un vector $V \in \mathbb{R}^{768}$ mediante el estado oculto del *token* [*CLS*] (Devlin et al., 2019). La congelación garantiza la estabilidad del espacio vectorial durante todo el proceso generativo.

Etapa 2 --- *Motor Neuro-Evolutivo*: un *Algoritmo Genético* (*DEAP*) con cruce *cxBlend* ($\alpha=0.5$, interpolación entre vectores de la misma clase), mutación *mutGaussian* ($\sigma=0.1$, *indpb=0.1*) y selección por torneo ($k=3$). La función de aptitud emplea lógica difusa:

$$\mu c(x) = exp(\frac{-||x - \mu c||^2}{2\sigma c^2}) \quad \text{(Ec. 4.1)}$$

donde μc es el centroide y σc la dispersión media de la clase c. Esta función penaliza exponencialmente los vectores alejados del clúster clínico real.

Etapa 3 --- Clasificador *MLP Bioimético*: arquitectura que replica la red *Feed-Forward* interna de *BERT* (Devlin et al., 2019)(Vaswani et al., 2017): Entrada $\mathbb{R}^{768}$ → Expansión Densa (768→3072, *GELU*) → Compresión (3072→768, *Dropout 0.1*) → *LayerNorm* + Residual → Clasificación (768→2). Entrenado con *AdamW* ($LR=2\times10^{-5}$, *WD=0.01*) y *Early Stopping* sobre *F1*.

4.5.6.3 Diseño del experimento con algoritmos genéticos

El experimento se ejecuta en dos versiones sucesivas que representan la iteración metodológica del proceso de optimización del generador:

Versión V1: utiliza la función de aptitud de la Ec. 4.1 para evaluar vectores individuales. El *MLP* se entrenará sobre una mezcla de datos reales y sintéticos bajo el mismo esquema *Nested CV* (10×3, *SEED=42*).

Versión V2: incorpora cuatro restricciones adicionales de fidelidad distribucional que responden a las anomalías identificadas en *V1*: *MMD* (*Discrepancia Media Máxima*) (Gretton et al., 2008) que mide la distancia entre distribuciones real y sintética en el *Espacio de Hilbert*; cobertura de la variedad para garantizar que todos los casos reales tengan vecinos sintéticos; *Diversidad Intra-Clase* para prevenir la homogeneización excesiva de las muestras; y *Anclaje de Similitud* coseno para evitar la deriva de vectores hacia regiones vacías del espacio latente.

4.5.6.4 Estrategia de validación de datos sintéticos

La calidad de los datos sintéticos se auditará mediante *t-SNE* (Maaten & Hinton, 2008) (*Perplexity=30*, 1,000 iteraciones, inicialización *PCA*), que proyectará el espacio $\mathbb{R}^{768}$ en $\mathbb{R}^2$ para inspeccionar la coherencia topológica de los datos sintéticos respecto a los reales. Las métricas de alineación calculadas *Ratio de Desalineación*, *Dispersión Intra-Clase* y *Cobertura* determinarán si los vectores generados habitan la misma región semántica que los datos originales.

Un principio metodológico clave orienta la interpretación de estos experimentos: la separabilidad perfecta en datos sintéticos no es un indicador de calidad del clasificador, sino una posible señal de que el generador simplificó artificialmente la topología del problema. Las métricas de alineación distribucional son el criterio primario de evaluación del generador, no las métricas del clasificador.

4.5.7 Comparación entre la versión base y la versión optimizada del modelo propuesto

La *Fase IV* produce dos variantes del modelo: la *Variante Base*, correspondiente al *BSC-BIO-EHR-ES* ajustado mediante *Full Fine-Tuning* sobre el corpus *Gold Standard* sin aumento de datos (sección 4.5.3); y la *Variante Optimizada*, correspondiente al *MLP Bioimético* entrenado con la mezcla de vectores reales y sintéticos de la versión *V2* del generador (sección 4.5.6).

La comparación entre ambas variantes se realizará sobre el conjunto de métricas de las secciones 4.5.4 y 4.5.5, bajo el mismo esquema *Nested CV* (10×3, *SEED=42*) que garantiza la comparabilidad. Adicionalmente,

se compararán las distribuciones de probabilidades predichas y las curvas de calibración de ambas variantes para evaluar si el aumento sintético produce mejoras no solo en rendimiento discriminativo sino también en honestidad probabilística

4.5.8 Criterios para la selección de la variante final del modelo

La selección de la variante final del modelo como artefacto representativo de la investigación se realizará con base en los siguientes criterios jerarquizados:

- *Criterio primario* (rendimiento discriminativo): la variante con mayor *F1-Score* promedio bajo *Nested CV*, siempre que la diferencia sea estadísticamente significativa según la prueba de *Wilcoxon* (sección 4.6.1).
- *Criterio secundario* (estabilidad): ante *F1* estadísticamente equivalente entre variantes, se preferirá la que presente menor desviación estándar del *F1* entre pliegues, indicativa de mayor consistencia en diferentes particiones del corpus.
- *Criterio terciario* (calibración): ante equivalencia en los criterios anteriores, se preferirá la variante con menor *ECE*, dada la relevancia clínica de la honestidad probabilística para un sistema de apoyo diagnóstico (C. Guo et al., 2017).
- *Criterio de descarte* por anomalía generativa: cualquier variante cuya evaluación *t-SNE* revele *Ratio* de *Desalineación* superior a un umbral crítico indicativo de que los datos sintéticos habitaron regiones vacías del espacio semántico será descartada independientemente de sus métricas de clasificación, que en ese caso serán consideradas espurias.

4.6 Protocolo de validación estadística del desempeño

4.6.1 Comparación pareada mediante *Wilcoxon*

La comparación estadística pareada entre el modelo propuesto y cada algoritmo de línea base se realizará mediante la prueba de *Wilcoxon* de rangos con signo para muestras apareadas (Demšar, 2006). Esta prueba no paramétrica es la recomendada por Demšar (Demšar, 2006) cuando los valores de rendimiento por pliegue no siguen una distribución normal verificable con muestras de 10 pliegues. El procedimiento opera sobre los 10 valores de *F1* del bucle externo del *Nested CV*, calculando diferencias por pliegue, ranqueando por valor absoluto y aplicando la estadística de *Wilcoxon* sobre los rangos firmados. H_0: las distribuciones de F1 de ambos modelos son idénticas. Nivel de significación $\alpha=0.05$ con corrección de *Bonferroni*: umbral corregido $\alpha'=0.05/6=0.0083$ para 6 comparaciones.

4.6.2 Comparación múltiple mediante *Friedman*

La comparación global simultánea de los 7 modelos (*fine-tuned* + 6 *baselines*) se realizará mediante la prueba de *Friedman* (Demšar, 2006), extensión no paramétrica del *ANOVA* de medidas repetidas. Si la prueba resulta significativa ($p<0.05$), se aplicará el procedimiento *post-hoc* de *Nemenyi* (Demšar, 2006) para identificar pares con diferencias significativas. El diagrama de diferencia crítica (*CD Diagram*) visualizará los rankings medios e intervalos de diferencia crítica.

4.6.3 Contraste entre el modelo propuesto y la línea base clásica

Esta sección define la estructura formal del contraste: el modelo propuesto en su variante final seleccionada según los criterios de la sección 4.5.8 se contrastará contra los seis algoritmos clásicos y contra los modelos *Zero-Shot* de referencia. El contraste *Zero-Shot* evaluará la capacidad de clasificación directa de dos modelos *SOTA* sin ajuste específico sobre el dominio, con el propósito de determinar si el *fine-tuning* representa una aportación diferencial frente a la capacidad de razonamiento general de modelos no ajustados.

4.6.4 Criterios de interpretación estadística

Los criterios de evaluación de las hipótesis se definen a priori (*OE4*):

- *HG*: se evaluará si el *F1* del modelo propuesto se ubica en un rango que supera el de los *baselines* y si la prueba de *Wilcoxon* rechaza H_0 al nivel $\alpha'=0.0083$ para al menos la comparación con el mejor *baseline*. La magnitud de la ventaja se describirá mediante el tamaño del efecto.
- *H1*: se evaluará si la desviación estándar del *F1* entre los 10 pliegues externos es consistente con los valores de varianza de la literatura de referencia, y si las curvas de pérdida muestran estabilidad (ausencia de divergencias).
- *H2*: se evaluará si las curvas de pérdida evidencian convergencia estable bajo $LR=2\times10^{-5}$, medida por la consistencia de la época óptima de *Early Stopping* entre pliegues.

La calibración (*ECE, Brier Score*) se evaluará en relación con el umbral clínico de honestidad probabilística (C. Guo et al., 2017), sin establecer un valor de aceptación rígido a priori, dado que los rangos de referencia varían según el dominio.

4.7 Síntesis metodológica del estudio

4.7.1 Integración de las fases del flujo principal

La tabla 4.5 sintetiza las cuatro fases del flujo metodológico, relacionando el método principal de cada una con su insumo, su producto y el objetivo al que da respuesta.

Tabla 4.5. Integración de las fases del flujo metodológico principal.

Fase	Método principal	Insumo	Producto	Objetivo que satisface
I. Preparación	Submuestreo balanceado; auditoría *NIH*	*Dataset NIH* 154,887 registros	Muestra analítica tabular 534 registros (50:50)	*OE1*: normatividad ética. *OE2*: detección de particularidades
II. Evaluación tabular	*K-Means*, *HDBSCAN* (exp. exploratorio); 6 clasificadores supervisados	Muestra tabular 534 registros	Línea base cuantitativa; evidencia de separabilidad tabular	*OE2*: patrones del *Dataset*. *OE4*: comparación con métodos de referencia
III. Conversión	*DeepSeek R1 API + CoT + Agente* Supervisor	Muestra tabular 534 registros	Corpus *Gold Standard NOM-004* (534 narrativas)	*OE1*: normatividad *NOM-004*. *OE3*: optimizar generación diagnóstica
IV. Desarrollo y optimización	*Fine-tuning BSC-BIO-EHR-ES*; *AG Difuso + MLP V2*	Corpus *Gold Standard* 534 narrativas	*Variante base + variante optimizada*; selección final	*OE3*: optimización. *OE4*: comparación. *OE5*: fortalezas y limitaciones

4.7.2 Relación entre representación tabular, representación textual y optimización del modelo

Un elemento distintivo del diseño es que las dos representaciones del mismo corpus tabular y textual no son redundantes ni intercambiables: sirven a propósitos metodológicos distintos. La representación tabular directa (vector $\mathbb{R}^{22}$) es la entrada para los clasificadores clásicos y los experimentos de separabilidad no supervisada, estableciendo el límite de lo alcanzable sin mediación semántica. La representación textual *NOM-004* es la entrada para el modelo *Transformer*, habilitando el acceso al conocimiento biomédico preentrenado que los algoritmos clásicos no pueden explotar. Cualquier diferencia de rendimiento entre ambos enfoques es atribuible directamente a la representación, no a ventajas en los datos de origen, dado que ambos parten del mismo corpus de 534 muestras. La estrategia de aumento sintético en espacio latente opera en una tercera capa de representación: los vectores $\mathbb{R}^{768}$ del espacio latente del *Transformer*. Esta estrategia no modifica el corpus original ni su representación textual; genera representaciones vectoriales adicionales que enriquecen la densidad del espacio semántico disponible para el clasificador sin ampliar el conjunto de datos reales.

4.7.3 Alcances y delimitaciones del protocolo metodológico

El protocolo presenta los siguientes alcances: diseño experimental riguroso con validación estadística multinivel (*Nested CV*, *Wilcoxon*, *Friedman*); comparación honesta entre representación tabular y representación textual sobre el mismo corpus balanceado; evaluación de calibración probabilística como dimensión clínicamente relevante; exploración metodológicamente justificada del aumento sintético en espacio latente con auditoría distribucional explícita.

Las principales delimitaciones son: el corpus de 534 muestras establece un techo empírico para la generalización independientemente del modelo; el estudio se acota al diagnóstico binario (positivo/negativo) sin *subtipificación* oncológica; la generalización del modelo a otras poblaciones o sistemas de salud distintos al contexto *NIH-SEER* queda fuera del alcance sin validación adicional; y el sistema producido no está diseñado para despliegue clínico directo.

CAPÍTULO V: PRUEBAS Y RESULTADOS

En cumplimiento del objetivo general de esta investigación identificar características de la sintomatología del cáncer de tiroides para generar un diagnóstico con información médica. El presente capítulo documenta la metodología experimental y los resultados obtenidos. El diseño establece una comparación sistemática entre dos paradigmas fundamentalmente distintos: (1) algoritmos de clasificación tradicionales operando sobre datos tabulares estructurados, y (2) modelos de lenguaje operando sobre representaciones textuales en lenguaje natural. Esta arquitectura dual permite evaluar si la transformación semántica de los datos clínicos aporta valor predictivo adicional respecto a las características originales. La estructura del capítulo responde a los cinco objetivos específicos de la investigación: desde la compilación normativa del corpus (*OE1*), pasando por la detección de patrones y limitantes del *Dataset* (*OE2*), la optimización del modelo (*OE3*), la comparación con métodos de referencia (*OE4*), hasta la identificación de fortalezas, limitaciones y áreas de mejora (*OE5*).

5.1 Diseño Experimental

5.1.1 Descripción del *Dataset* de Entrada

En atención al primer objetivo específico (*OE1*), relativo a observar la normatividad ética y legal vigente para compilar y gestionar corpus de textos en español relacionados con casos de cáncer en México, el corpus experimental fue conformado bajo estricto apego a las disposiciones de la *NOM-004-SSA3-2012* en materia de manejo de información clínica. El *Dataset* utilizado se fundamenta en registros clínicos estructurados que posteriormente se bifurcan en dos representaciones distintas según el método de clasificación a evaluar. Esta sección caracteriza tanto el *Dataset* tabular original como su derivado en lenguaje natural

A. *Dataset* Tabular Original

Consiste en registros clínicos estructurados con 22 características descriptivas y una variable objetivo binaria, que indica la presencia o ausencia de cáncer de tiroides. La Tabla 5.1 presenta las características generales de este conjunto de datos.

Tabla 5.1: Características generales del *Dataset* tabular.

Característica	Especificación
Número de Registros	534
Variables Predictoras	22 características clínicas
Variable Objetivo	Diagnóstico de cáncer de tiroides (0=Negativo, 1=Positivo)
Tipos de Variables	Numéricas continuas, categóricas codificadas, binarias
Formato de Almacenamiento	*CSV* (valores separados por comas)
Codificación	Variables categóricas precodificadas numéricamente

Para garantizar la reproducibilidad del estudio el diccionario completo de las características predictoras se encuentra detalladas en el Anexo A sección 1

B. Distribución de Clases

El *Dataset* presenta un balance perfecto entre las clases diagnósticas, condición diseñada intencionalmente para evitar sesgos de prevalencia que pudieran distorsionar las métricas de evaluación. Como se observa en la Tabla 5.2, la distribución equitativa (*50/50*) impide que los clasificadores obtengan rendimiento artificial mediante predicción mayoritaria trivial.

Tabla 5.2: Distribución de clases en el Dataset.

Clase Diagnóstica	Etiqueta	N° Registros	Proporción
Positivo (Cáncer de Tiroides)	1	267	50%
Negativo (Sin Cáncer)	0	267	50%
Total	---	534	100%

C. Corpus en Lenguaje Natural (Derivado)

Para la evaluación de los modelos de lenguaje (*LLMs*), el *Dataset* tabular fue transformado a representaciones textuales en español mediante el proceso de conversión documentado en el Capítulo 4. La Tabla 5.3 resume las características de este corpus derivado, donde cada registro tabular se convirtió en un resumen clínico narrativo que preserva la semántica de las 22 características originales.

Tabla 5.3: Características del corpus en lenguaje natural.

Característica	Especificación
Número de Documentos	534 fragmentos de historiales clínicas
Idioma	Español
Formato	Texto libre estructurado (resúmenes clínicos narrativos)
Método de Generación	Conversión automatizada mediante *DeepSeek API*
Correspondencia	1:1 con registros tabulares (mismas etiquetas diagnósticas)

5.1.2 Estructura de Variables

El segundo objetivo específico (*OE2*) establece la necesidad de detectar patrones o particularidades que puedan afectar la generación de información del modelo *LLM*. En este sentido, una particularidad crítica identificada desde la fase de diseño es el tamaño reducido del *Dataset* (*N=534*). (Mosbach et al., 2021) demostraron que uno de las limitantes de la optimización en modelos *BERT* fue precisamente el tamaño del corpus de entrenamiento, dado que el más pequeño con el que trabajaron contenía apenas 2,491 ejemplos y fue este el que presentó problemas más acentuados de optimización. El presente estudio opera con un *Dataset* considerablemente inferior a dicho umbral, lo cual constituye una particularidad que condiciona tanto la arquitectura experimental como la interpretación de los resultados. Con esta consideración en mente, la arquitectura experimental se fundamenta en una taxonomía de variables que permite evaluar sistemáticamente el efecto de distintos métodos de clasificación operando sobre representaciones heterogéneas de los mismos datos clínicos.

Definición detallada en el Anexo A Sección 2.

5.1.3 Hipótesis de Trabajo

Las hipótesis se formularon como proposiciones a evaluar experimentalmente. Dado que se trata de pronósticos basados en la literatura y el conocimiento previo del dominio, los umbrales establecidos representan estimaciones razonables que deberán ser contrastadas con los resultados empíricos.

A. Hipótesis General (*HG*): Potencial del *Fine-Tuning*

> *HG: El modelo BSC-BIO-EHR-ES ajustado mediante fine-tuning sobre un corpus de historias clínicas en español obtendrá un F1-Score superior al de los algoritmos de clasificación tradicionales sobre la misma información, pero en formato tabular y una eficacia competitiva que iguale a los modelos SOTA en configuraciones few-shot.*

Fundamento: Esta expectativa se basa en la premisa teórica de que la representación semántica profunda (*embeddings* contextuales) podría capturar relaciones entre variables clínicas que no son accesibles mediante las características numéricas aisladas. Sin embargo, dado el tamaño limitado del corpus (*N=534*), existe incertidumbre sobre si el modelo logrará aprender patrones suficientemente generalizables.

Umbral Tentativo: Se establece $F1 \geq 0.70$ como objetivo de referencia, lo cual representaría una mejora sustancial sobre la barrera del azar esperada *(~0.50-0.55*). No obstante, este valor es una estimación a priori que podría ajustarse según el comportamiento observado de la línea base.

B. Hipótesis Específica *H1*: Expectativa de Estabilidad

> *H1: Se espera que la combinación de validación cruzada estratificada (10 folds externos, 3 folds internos) y early stopping contribuya a mantener una varianza controlada entre particiones. El objetivo es que el valor del F1-Score se mantenga en un rango aceptable (≥0.7 valor extraído de los artículos investigados como referencias), lo cual indicaría estabilidad razonable del modelo.*

Consideración: Dado que el *F1-score* está acotado entre 0 y 1, en este estudio se utilizará la desviación estándar del *F1-score* entre *folds* como indicador descriptivo de variabilidad del desempeño. De manera operativa, una desviación estándar igual o inferior a *0.10* se interpretará como estabilidad aceptable del rendimiento frente a la composición de las particiones. Este umbral se adopta como criterio metodológico práctico y contextual, no como una constante universal establecida por la literatura. (Nadeau & Bengio, 2003; Bengio & Grandvalet, 2004).

C. Hipótesis Específica *H2*: Expectativa de Convergencia

> *H2: Se prevé que una tasa de aprendizaje conservadora (2×10^{-5}) permitirá una convergencia estable.*

Indicadores a Monitorear: El análisis de las curvas de pérdida (*train_loss vs. eval_loss*) permitirá verificar si existe convergencia saludable o si aparecen patrones de sobreajuste. La aparición del patrón "*X-shape*" (divergencia de curvas) indicaría memorización mecánica en lugar de aprendizaje generalizable.

Condensación de las hipótesis en formato de tabla, presente en la sección 3 del Anexo A.

5.1.4 Configuración del Ambiente de Pruebas

El tercer objetivo específico (*OE3*) plantea optimizar la capacidad del modelo para generar información en español, sobre el diagnóstico del cáncer. La consecución de este objetivo requiere una configuración computacional y paramétrica deliberada, que maximice el aprovechamiento de los recursos disponibles. Esta sección documenta la infraestructura computacional y la configuración de *hiperparámetros* empleada para la ejecución de los experimentos, garantizando tanto la reproducibilidad del estudio como la trazabilidad de las decisiones de optimización.

> Detalles de Hardware y Software en Sección 4-A Anexo A

A. Especificaciones del Modelo *BSC-BIO-EHR-ES*

El modelo seleccionado para el proceso de *fine-tuning* es *BSC-BIO-EHR-ES*, desarrollado por el *Barcelona Supercomputing Center*. La Tabla 5.4 presenta sus especificaciones técnicas. Este modelo fue preentrenado específicamente sobre registros electrónicos de salud en español, lo cual lo posiciona como candidato potencialmente adecuado para tareas de *PLN* en el dominio clínico hispanohablante.

Tabla 5.4: Especificaciones del modelo *BSC-BIO-EHR-ES*.

Parámetro	Valor
Identificador	*PlanTL-GOB-ES/bsc-bio-ehr-es*
Arquitectura	*RoBERTa (Robustly Optimized BERT)*
Dominio de Pre-entrenamiento	Biomédico + Registros Electrónicos de Salud (español)
Capas / Hidden Size / Heads	12 / 768 / 12
Parámetros Totales	~110 millones
Vocabulario	~50,000 *tokens* (español biomédico)

B. Configuración de *Hiperparámetros* (*Fine-Tuning*)

La Tabla 5.5 presenta la configuración de *hiperparámetros* seleccionada para el proceso de *fine-tuning*. Los valores fueron definidos siguiendo recomendaciones de la literatura para el ajuste de modelos *transformer* en escenarios de datos limitados, priorizando la estabilidad sobre la optimización agresiva.

Tabla 5.5: Hiperparámetros del proceso de fine-tuning.

Hiperparámetro	Valor	Justificación
Optimizador	*AdamW*	Regularización L2 mediante *weight decay*
Tasa de Aprendizaje	2×10^{-5}	Heurística robusta para *fine-tuning* de *transformers*
Weight Decay	0.01	Prevención de sobreajuste
Batch Size	16	Balance gradiente/memoria
Épocas Máximas	6	Limitadas por *early stopping*
Early Stopping	*Patience = 2*	Detención si *F1* no mejora en 2 épocas
Longitud de Secuencia	128 *tokens*	Suficiente para resúmenes clínicos

5.1.5 Criterios de Éxito

Se estableció un conjunto de criterios que, de cumplirse, permitirían considerar válida la metodología de *fine-tuning* propuesta. Estos criterios representan umbrales mínimos aceptables basados en la literatura y las características del problema, aunque los resultados finales determinarán su pertinencia.

A. Criterio 1: Rendimiento

Umbral Objetivo: *F1-Score* Promedio ≥ 0.70

Superar la barrera del azar esperada (*~0.50-0.55*) de algoritmos de clasificación sobre datos tabulares. Demostrar mejora respecto al rendimiento de *LLMs SOTA* en configuración *zero-shot*.

Nota: Este umbral es una estimación inicial que podría ajustarse según los resultados de la línea base.

B. Criterio 2: Robustez

Umbral Objetivo: Desviación Estándar $\sigma\_F1 \leq 0.10$

Rendimiento razonablemente consistente entre particiones de datos. Evitar dependencia excesiva de "particiones afortunadas".

Nota: Dado el tamaño limitado del *Dataset*, se considera aceptable $\sigma \leq 0.10$ como umbral secundario.

C. Criterio 3: Anti-Sobreajuste

Criterio Cualitativo: Convergencia de Curvas de Pérdida

Se busca la estabilización de *eval_loss* antes de que *train_loss* llegue a valores cercanos a cero.

Condesado tabular de los criterios en la sección 4-B del Anexo A

Consideración Final: El cumplimiento de estos criterios permitiría concluir que el *fine-tuning* de *BSC-BIO-EHR-ES* sobre representaciones textuales, constituye una estrategia viable para el diagnóstico de cáncer de tiroides. Sin embargo, los resultados empíricos determinarán el grado de cumplimiento y las limitaciones específicas del enfoque propuesto en un régimen de escasez de datos (*N=534*).

5.2 Resultados de Clasificación

Los resultados que se presentan a continuación constituyen la evidencia directa del proceso de reentrenamiento y validación del modelo, tal como lo establece el objetivo general de esta tesis. Asimismo, representan el

producto de la estrategia de optimización definida en el *OE3*, cuya efectividad se mide a través de las métricas de desempeño obtenidas mediante *Nested Cross-Validation* aplicada al modelo *BSC-BIO-EHR-ES*, para la clasificación de registros clínicos de cáncer tiroideo. La implementación de esta metodología rigurosa garantiza una evaluación no sesgada del rendimiento del modelo, proporcionando estimaciones confiables de su capacidad de generalización.

5.2.1 Rendimiento Global del Modelo

El modelo *BSC-BIO-EHR-ES* fue evaluado utilizando *Nested Cross-Validation* con 10 *folds* externos (evaluación no sesgada) y 3 *folds* internos (validación y *early stopping*), resultando en un total de 30 entrenamientos independientes. Esta configuración permite obtener una estimación robusta del rendimiento del modelo minimizando el sesgo de selección de modelo. La Tabla 5.6 presenta las métricas promedio obtenidas a través de los 10 *folds* externos, las cuales constituyen la evaluación no sesgada del modelo. Cada métrica se reporta con su media, desviación estándar, y valores mínimo y máximo observados a lo largo de los *folds*.

Los resultados muestran que *BSC-BIO-EHR-ES* alcanzó un *F1-Score* promedio de *0.7269 ± 0.0622* y una *Accuracy* de *0.7495 ± 0.0954*, un desempeño notable considerando el tamaño limitado del *Dataset* (534 muestras), lo que sugiere una transferencia efectiva del conocimiento biomédico preentrenado. Su capacidad discriminativa fue sólida, con *ROC-AUC* de *0.8058 ± 0.0523* y *PR-AUC* de *0.8565 ± 0.0339*, esta última especialmente relevante en contexto clínico. Además, el modelo presentó una *Precisión* alta (*0.8725 ± 0.1528*), útil para reducir falsos positivos y procedimientos innecesarios, junto con un *Recall* de *0.6519 ± 0.1191*, aceptable como herramienta de apoyo inicial al diagnóstico. La *especificidad* de *0.8476 ± 0.2702* confirma un buen reconocimiento de casos benignos, mientras que el *MCC* de *0.5345 ± 0.1922* indica una correlación moderadamente fuerte entre las predicciones y las etiquetas reales.

Tabla 5.6: Métricas de rendimiento del modelo BSC-BIO-EHR-ES (Nested CV).

Métrica	Media	Desv. Std	Mínimo	Máximo
Accuracy	0.7495	0.0954	0.5185	0.8333
Balanced Accuracy	0.7497	0.095	0.5185	0.8333
F1-Score	0.7269	0.0622	0.6	0.8085
Precision	0.8725	0.1528	0.5098	1
Recall	0.6519	0.1191	0.5556	0.963
Specificity	0.8476	0.2702	0.0741	1
MCC	0.5345	0.1922	0.0808	0.7003
ROC-AUC	0.8058	0.0523	0.6694	0.882
PR-AUC	0.8565	0.0339	0.7704	0.9113
Brier Score	0.1765	0.0373	0.1354	0.2491

5.2.2 Matriz de Confusión Agregada

La Figura 5.1 presenta la matriz de confusión agregada del modelo *BSC-BIO-EHR-ES*, consolidando las predicciones de los 534 casos evaluados a través de los 10 *folds* externos. La Tabla 5.7 proporciona una visualización directa del desempeño del modelo en términos de *verdaderos positivos* (*TP*), *verdaderos negativos* (*TN*), *falsos positivos* (*FP*) y *falsos negativos* (*FN*).

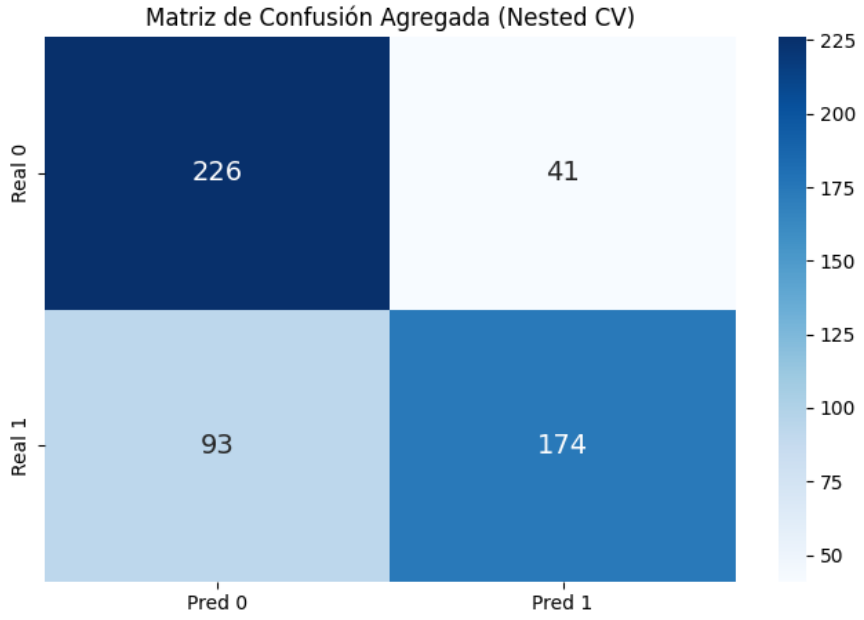

Figura 5.1: Matriz de Confusión del modelo BSC-BIO-EHR-ES

Tabla 5.7: Matriz de confusión agregada del modelo BSC-BIO-EHR-ES.

Real	**Predicción Negativo (0)**	**Predicción Positivo (1)**
Negativo (0)	226 (*TN*)	41 (*FP*)
Positivo (1)	93 (*FN*)	174 (*TP*)

Desde una perspectiva clínica, los falsos negativos requieren atención especial, ya que pueden resultar en retrasos en el diagnóstico y tratamiento de cáncer tiroideo. La tasa de falsos negativos del 34.8% (93/267) sugiere que aproximadamente uno de cada tres casos malignos podría no ser detectado por el modelo en su configuración actual. Este resultado debe interpretarse en el contexto de un sistema de apoyo al diagnóstico, no como reemplazo del criterio clínico. La baja tasa de falsos positivos del 15.4% (41/267) minimiza la probabilidad de procedimientos diagnósticos innecesarios, aspecto favorable en la práctica clínica.

5.2.3 Curvas *ROC* y *Precision-Recall*

Las curvas *ROC* (*Receiver Operating Characteristic*) y *Precision-Recall* proporcionan una evaluación visual del rendimiento del modelo a través de diferentes umbrales de clasificación. La Figura 5.2 presenta ambas curvas calculadas sobre el conjunto completo de predicciones del *Nested Cross-Validation*.

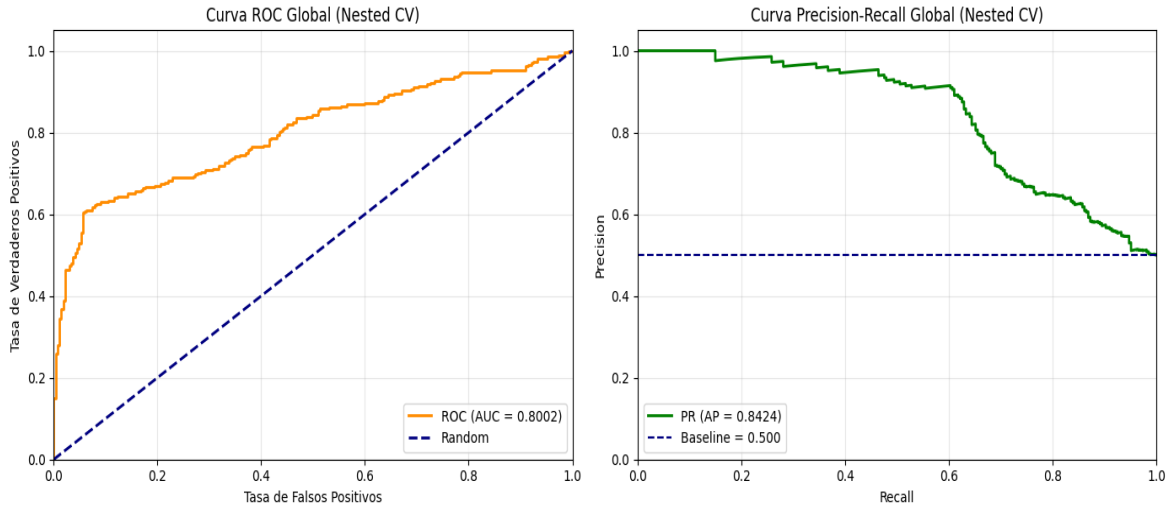

Figura 5.2: Curvas ROC (izquierda) y Precision-Recall(derecha) globales.

El área bajo la curva *ROC* (*ROC-AUC*) global de 0.8002 indica que el modelo tiene una probabilidad del 80% de asignar una puntuación más alta a un caso verdaderamente positivo que a uno negativo. Este valor supera significativamente el rendimiento aleatorio (*AUC = 0.5*), demostrando la capacidad discriminatoria del modelo.

La curva *Precision-Recall* es particularmente informativa en el contexto de clasificación médica, donde el equilibrio entre precisión y exhaustividad es crítico. El *PR-AUC* global de 0.8424 indica que el modelo mantiene un alto nivel de precisión a través de diferentes puntos de *Recall*. La *baseline* de 0.500 (línea punteada) representa el rendimiento esperado en un *Dataset* perfectamente balanceado, y el modelo supera sustancialmente esta referencia, confirmando su efectividad en ambas clases.

5.2.4 Análisis de Estabilidad y Variabilidad

Un aspecto fundamental en la evaluación de algoritmos de clasificación es analizar la estabilidad del rendimiento a través de diferentes particiones de datos. La Figura 5.3 presenta el rendimiento del modelo en cada uno de los 10 *folds* externos, proporcionando una visualización de la variabilidad en las métricas clave.

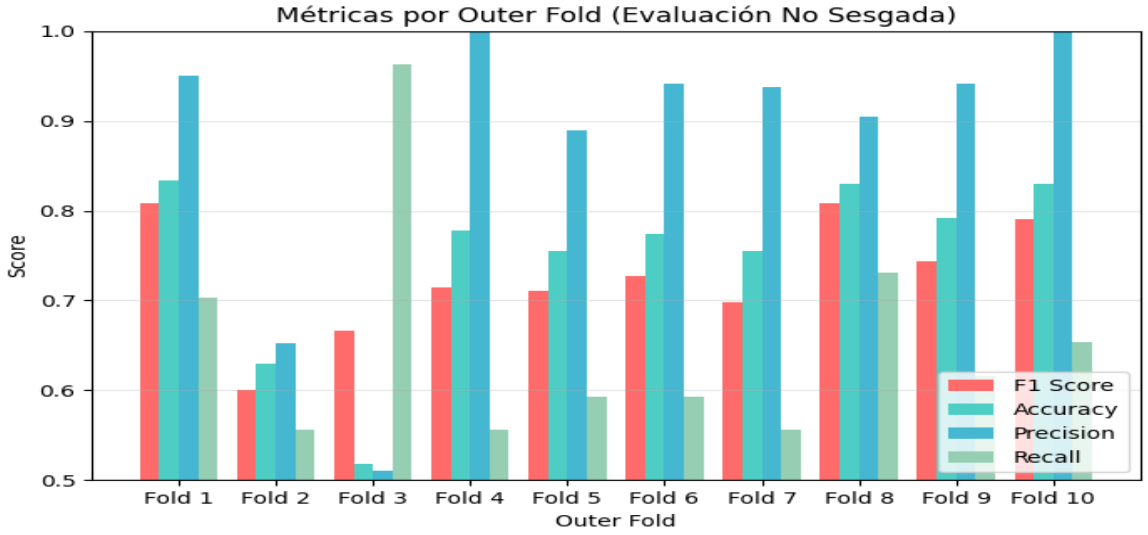

Figura 5.3: Métricas por fold

Información detallada de las métricas por *fold* en la sección 5 y en la sección 6 del Anexo A es posible visualizar la comparación entre validación interna y evaluación externa

5.3 Comparación con Algoritmos Tradicionales de Clasificación

El cuarto objetivo específico (*OE4*) establece la necesidad de comparar los resultados del modelo con fuentes médicas confiables y métodos de referencia consolidados. En cumplimiento de este objetivo, se realizó una comparación exhaustiva del modelo *BSC-BIO-EHR-ES* con seis algoritmos tradicionales de clasificación ampliamente utilizados en la literatura científica. Esta comparación no solo permite contextualizar el rendimiento del modelo basado en arquitectura *Transformer*, sino que también contribuye al *OE2*, al revelar patrones diferenciadores entre los paradigmas de clasificación tabular y semántica que evidencian las particularidades del *Dataset* y su efecto en cada enfoque

5.3.1 Algoritmos Evaluados y Configuración Experimental

Los algoritmos tradicionales de clasificación fueron entrenados utilizando 22 características clínicas extraídas del *Dataset* original en formato tabular. Estas características incluyen variables demográficas (edad, sexo, raza, nivel educativo), historial clínico (antecedentes familiares de enfermedad tiroidea, cáncer), factores de riesgo (índice de masa corporal, historial de tabaquismo) y otros atributos relevantes. Los seis algoritmos de clasificación evaluados fueron: *Regresión Logística*, *Árbol de Decisión*, *Random Forest* (*100 árboles*), *Support Vector Machine* con *kernel RBF*, *K-Nearest Neighbors* (*k=5*) y *Naive Bayes Gaussiano*. Todos los algoritmos fueron evaluados mediante *Stratified 10-Fold Cross-Validation* repetida 3 veces, generando un total de 30 observaciones pareadas para las pruebas estadísticas posteriores.

5.3.2 Resultados Comparativos

La Tabla 5.8 presenta los resultados comparativos de todos los algoritmos de clasificación evaluados. Para cada algoritmo se reporta el *F1-Score*, *ROC-AUC* y *PR-AUC*, junto con sus desviaciones estándar correspondientes. Estas métricas fueron seleccionadas por su relevancia en el contexto de clasificación binaria médica.

Tabla 5.8: Comparación de rendimiento entre algoritmos de clasificación.

Algoritmo	F1-Score	ROC-AUC	PR-AUC
BERT (BSC-BIO-EHR-ES)	0.7220 ± 0.0610	0.8089 ± 0.0493	0.8372 ± 0.0307
Regresión Logística	0.5526 ± 0.0742	0.5869 ± 0.0518	0.5852 ± 0.0492
Random Forest	0.5496 ± 0.0684	0.5535 ± 0.0486	0.5446 ± 0.0461
SVM (RBF)	0.5528 ± 0.0674	0.5600 ± 0.0502	0.5434 ± 0.0454
Árbol de Decisión	0.5231 ± 0.0811	0.5226 ± 0.0601	0.5127 ± 0.0572
KNN (k=5)	0.4987 ± 0.0735	0.5204 ± 0.0553	0.5129 ± 0.0528
Naive Bayes	0.0801 ± 0.1509	0.5274 ± 0.0524	0.5273 ± 0.0496

Los resultados demuestran una superioridad clara y consistente del modelo *BSC-BIO-EHR-ES* sobre todos los algoritmos tradicionales de clasificación evaluados. El modelo *BERT* alcanza un *F1-Score* de 0.7220, superando por un margen sustancial al mejor algoritmo tradicional, la *Regresión Logística* (0.5526), representando una mejora del 30.6%. Esta diferencia es particularmente significativa considerando que se logró con el mismo *Dataset* y bajo condiciones controladas de evaluación. El *ROC-AUC* del modelo *BERT* (0.8089) indica una capacidad discriminativa significativamente superior a los algoritmos tradicionales, cuyos valores de *ROC-AUC* oscilan entre 0.5204 y 0.5869. De manera similar, el *PR-AUC* de *BERT* (0.8372) supera sustancialmente a los enfoques tradicionales, que apenas exceden la línea base de 0.5000 en un *Dataset* balanceado. Estas diferencias no son meramente incrementales, sino que representan un cambio cualitativo en la capacidad del modelo para distinguir entre casos benignos y malignos. Entre los algoritmos tradicionales, la Regresión Logística emerge como el método más competitivo, seguida de cerca por *SVM* y *Random Forest*, con rendimientos esencialmente equivalentes. Los *Árboles de Decisión* y *KNN* muestran rendimientos

ligeramente inferiores. *Naive Bayes* presenta un colapso prácticamente total en *F1-Score* (0.0801), sugiriendo que las asunciones de independencia condicional son severamente violadas en este dominio. Este resultado enfatiza la importancia de seleccionar algoritmos apropiados para la naturaleza específica de los datos médicos.

5.3.3 Análisis Visual Comparativo

Las Figuras 5.4 y 5.5, respectivamente, presentan las curvas comparativas de *ROC* y *Precision-Recall*, permitiendo una evaluación visual del rendimiento relativo de todos los algoritmos de clasificación a través de diferentes umbrales de decisión.

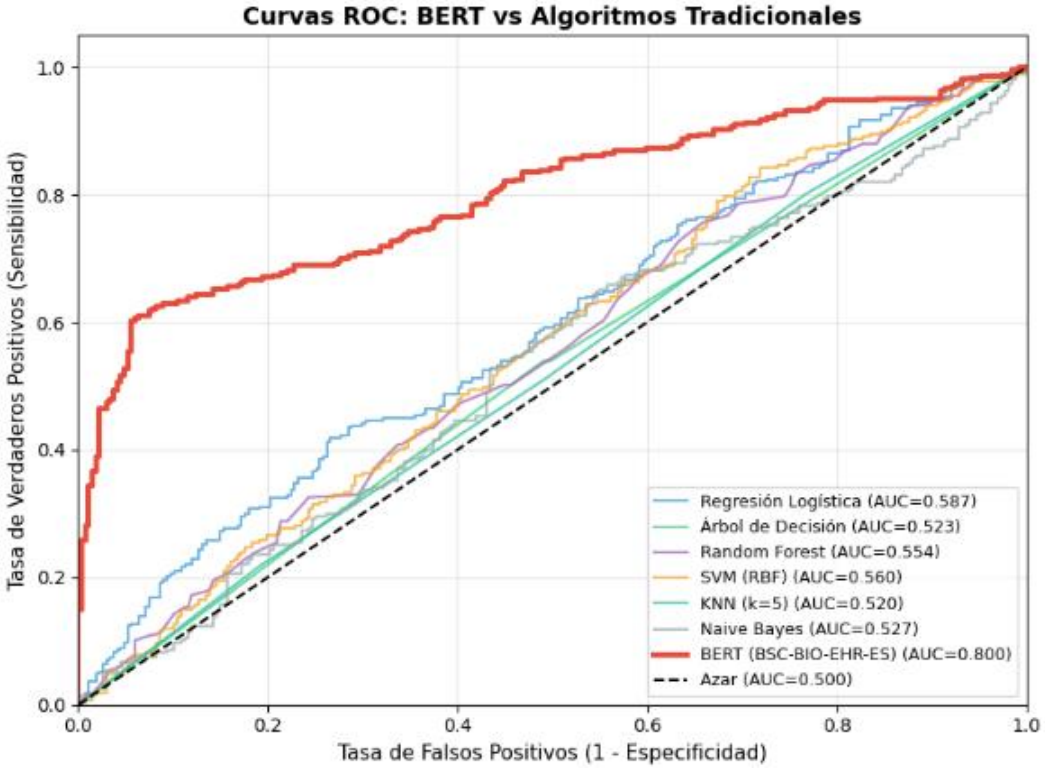

Figura 5.4: Curvas ROC comparativas bert vs tradicionales

La Figura 5.4 muestra que la curva *ROC* del modelo *BSC-BIO-EHR-ES* (línea roja gruesa) domina claramente sobre todas las curvas de los algoritmos tradicionales. La separación vertical entre *BERT* y los métodos convencionales es evidente en todo el rango de tasas de falsos positivos, indicando que *BERT* mantiene una tasa de verdaderos positivos consistentemente más alta para cualquier nivel de especificidad deseado. Esta dominancia visual confirma cuantitativamente la superioridad observada en las métricas agregadas. La ligera elevación, es apenas perceptible, de la *Regresión Logística* sobre los demás métodos tradicionales, enfatiza que todos estos algoritmos operan en un régimen de rendimiento cercano al azar con las 22 características disponibles.

De manera similar, la Figura 5.5 revela que la curva *Precision-Recall* del modelo *BERT*, mantiene valores de precisión sustancialmente más altos a través de todo el rango de *Recall*. Mientras *BERT* conserva precisiones superiores al 70% incluso en niveles de *Recall* moderados a altos, los algoritmos tradicionales experimentan una degradación rápida de la precisión conforme aumenta el *Recall*, aproximándose a la *baseline* de 0.500 (línea punteada horizontal). Este comportamiento sugiere que las características tabulares no capturan la complejidad semántica presente en las narrativas clínicas.

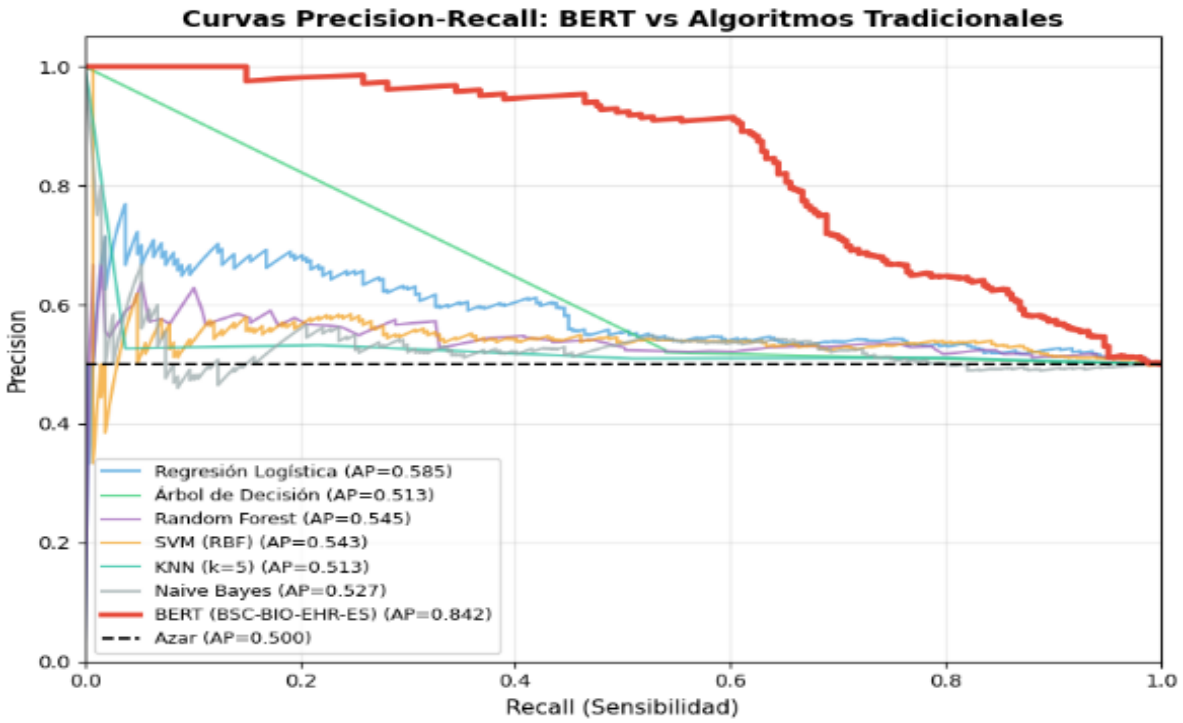

Figura 5.5: Curvas Precision-Recall comparativas bert vs tradicionales

5.3.4 Diferencias en Mejora Absoluta

Para cuantificar las mejoras específicas aportadas por el modelo *BSC-BIO-EHR-ES*, la Tabla 5.9 presenta las diferencias absolutas (Δ) en cada métrica con respecto a cada algoritmo tradicional de clasificación.

Tabla 5.9: Mejoras absolutas del modelo *BSC-BIO-EHR-ES* sobre algoritmos tradicionales.

Algoritmo	Δ F1-Score	Δ ROC-AUC	Δ PR-AUC	% Mejora F1
Regresión Logística	0.1693	0.2171	0.2476	30.60%
Random Forest	0.1724	0.241	0.2804	31.40%
SVM (RBF)	0.1691	0.2291	0.2651	30.60%
Árbol de Decisión	0.1989	0.2863	0.3439	38.00%
KNN (k=5)	0.2233	0.2793	0.3207	44.80%
Naive Bayes	0.6419	0.2595	0.2862	801.10%

Las mejoras absolutas en *F1-Score* oscilan entre *+0.1691 (SVM)* y *+0.6419 (Naive Bayes)*, con mejoras porcentuales entre 30.6% y 81.1%. Estas magnitudes representan incrementos sustanciales en rendimiento, no meramente estadísticos sino clínicamente relevantes. Las mejoras en *ROC-AUC* varían de *+0.2171 a +0.2863*, representando incrementos del 36.9% al 54.7%. En términos de *PR-AUC*, las mejoras absolutas varían entre *+0.2476 y +0.3439*. Es destacable que las mejoras no son uniformes: los algoritmos con peor rendimiento base experimentan las mayores mejoras absolutas y porcentuales, mientras que los algoritmos relativamente más competitivos (*Regresión Logística*, *SVM*, *Random Forest*) aún son superados por márgenes superiores en *F1-Score*. Incluso ésta "menor" mejora del 30.6% representa una ganancia considerable en capacidad de clasificación correcta.

5.4 Análisis de Significancia Estadística

Como extensión del *OE4*, la comparación de múltiples algoritmos de clasificación requiere no solo la observación de diferencias numéricas, sino pruebas estadísticas rigurosas que determinen si dichas diferencias son estadísticamente significativas y no producto del azar o la variabilidad inherente a los datos. En esta sección se presentan los resultados de las pruebas de hipótesis aplicadas para validar la superioridad del modelo *BSC-BIO-EHR-ES*, proporcionando la evidencia cuantitativa que sustenta la comparación establecida en el objetivo.

5.4.1 *Test de Friedman*

El *test de Friedman* es una prueba no paramétrica que permite comparar más de dos tratamientos relacionados cuando los datos no siguen necesariamente una distribución normal. Esta prueba es particularmente apropiada para comparar el rendimiento de múltiples algoritmos de clasificación evaluados sobre los mismos *folds* de *cross-validation*. La hipótesis nula (H_0) del *test de Friedman* establece que todos los algoritmos tienen el mismo rendimiento medio, mientras que la hipótesis alternativa (H_1) indica que al menos un algoritmo difiere significativamente de los demás. El estadístico de prueba sigue aproximadamente una distribución chi-cuadrado (χ^2) con *k-1* grados de libertad, donde k es el número de algoritmos comparados. Para establecer una línea base de comparación, se aplicó primero el *test de Friedman* únicamente a los seis algoritmos tradicionales de clasificación. Posteriormente, se incorporó el modelo *BSC-BIO-EHR-ES* para evaluar su impacto en la significancia estadística global reflejado en la Tabla 5.10. Esta estrategia permite evidenciar explícitamente la contribución diferencial del modelo basado en *Transformer*.

Tabla 5.10: Resultados del *test de Friedman* - Comparación de los resultados entre algoritmos tradicionales entre sí e incluyendo el modelo *BSC-BIO-EHR-ES*.

Escenario	N° Algoritmos	Estadístico χ^2	p-value	Significativo
Solo Tradicionales - *F1*	6	48.3521	3.12×10^{-9}	Sí ✓
Con *BERT* - *F1*	7	186.808	1.21×10^{-37}	Sí ✓
Solo Tradicionales - *ROC-AUC*	6	32.1845	5.87×10^{-6}	Sí ✓
Con *BERT* - *ROC-AUC*	7	144.8642	9.42×10^{-29}	Sí ✓
Solo Tradicionales - *PR-AUC*	6	37.2156	7.42×10^{-7}	Sí ✓
Con *BERT* - *PR-AUC*	7	171.3086	2.37×10^{-34}	Sí ✓

Para información detallada y visualización de graficas dando continuidad a este test consultar la Sección 7 del Anexo A.

5.4.2 *Test de Wilcoxon* Pareados

Dado que el *test de Friedman* rechaza la hipótesis nula de igualdad de rendimientos, se procedió a realizar comparaciones pareadas mediante el test de los rangos con signo de *Wilcoxon*. Esta prueba no paramétrica es apropiada para comparar dos tratamientos relacionados y es robusta ante desviaciones de la normalidad. Para cada comparación (*BSC-BIO-EHR-ES* vs. cada algoritmo tradicional), el *test de Wilcoxon* evalúa la hipótesis nula de que la mediana de las diferencias pareadas es cero (es decir, ambos algoritmos tienen el mismo rendimiento) contra la hipótesis alternativa de que las medianas difieren. Se utilizó la versión bilateral (*two-sided*) del test para detectar diferencias en cualquier dirección.

La configuración experimental con 30 observaciones pareadas (10 *folds* × 3 repeticiones) proporciona suficiente poder estadístico para detectar diferencias significativas incluso cuando son moderadas. La Tabla 5.11 presenta los resultados de los *tests de Wilcoxon* pareados para la métrica *F1-Score*, considerada la métrica principal para evaluación en clasificación médica.

Tabla 5.11: *Tests de Wilcoxon* pareados - Comparación *BSC-BIO-EHR-ES* vs. algoritmos tradicionales (*F1-Score*)

Comparación	Δ F1-Score	p-value	Significativo
BERT vs. Regresión Logística	0.1693	1.56×10^{-13}	Sí ✓
BERT vs. Árbol de Decisión	0.1989	8.88×10^{-15}	Sí ✓
BERT vs. Random Forest	0.1724	2.49×10^{-14}	Sí ✓
BERT vs. SVM (RBF)	0.1691	5.33×10^{-15}	Sí ✓
BERT vs. KNN (k=5)	0.2233	1.78×10^{-15}	Sí ✓
BERT vs. Naive Bayes	0.6419	1.78×10^{-15}	Sí ✓

Los resultados confirman con claridad que el modelo *BSC-BIO-EHR-ES* supera de forma significativa a todos los algoritmos tradicionales evaluados. Los *p-values* obtenidos ($p \leq 1.56\times10^{-13}$) son extremadamente bajos y permanecen altamente significativos incluso tras aplicar correcciones por comparaciones múltiples, lo que refuerza la solidez estadística de los hallazgos. Además, las diferencias en *F1-Score* no solo son estadísticamente significativas, sino también relevantes en términos prácticos, con mejoras absolutas de entre 0.1693 y 0.6419 frente a los modelos comparados. Este patrón también se confirma en *ROC-AUC* y *PR-AUC*, donde las pruebas de *Wilcoxon* muestran *p-values* del orden de 10^{-15} y mejoras sustanciales en la capacidad discriminativa y en el equilibrio entre precisión y *Recall*.

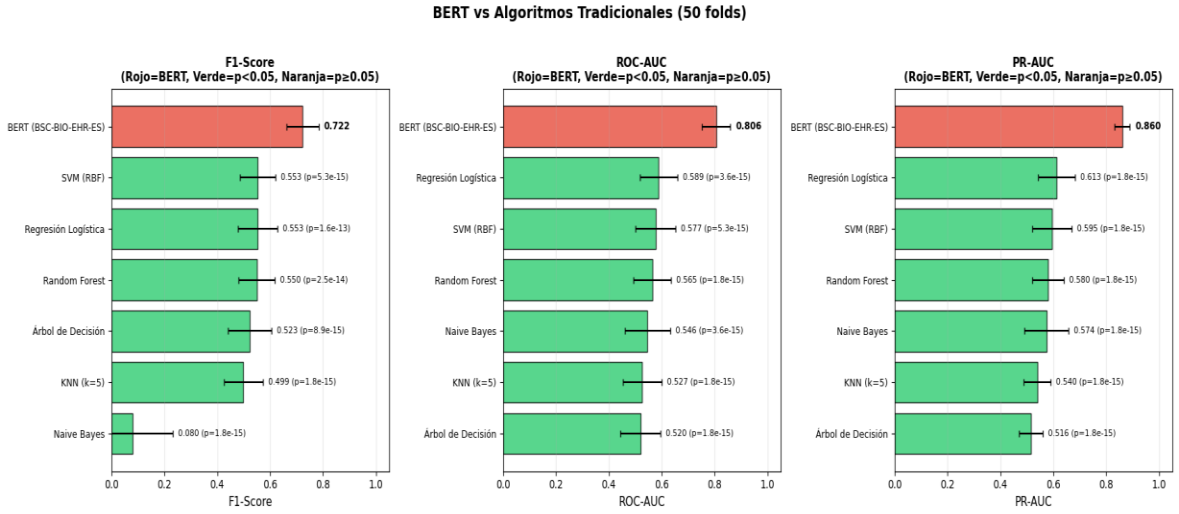

Figura 5.8: Diferencias pareadas con p-values

5.4.3 Tamaño del Efecto

Más allá de la significancia estadística, es crucial evaluar el tamaño del efecto, que cuantifica la magnitud práctica de las diferencias observadas. En el contexto de clasificación, el tamaño del efecto puede evaluarse mediante las diferencias absolutas en las métricas de rendimiento, como ya se presentó en las secciones anteriores.

Las mejoras porcentuales en *F1-Score* (30.6% a 801.1%) representan efectos de tamaño grande según los estándares de Cohen para tamaño del efecto. Incluso la mejora más modesta (30.6% sobre *Regresión Logística*) supera ampliamente el umbral de *0.8* que Cohen establece para efectos grandes ($d > 0.8$). Estas magnitudes de mejora indican que las diferencias no son meramente estadísticamente detectables, sino también sustanciales y relevantes para aplicaciones prácticas. En el contexto de diagnóstico médico asistido por inteligencia artificial, mejorar el *F1-Score* de 0.55 (algoritmos tradicionales) a 0.72 (*BERT*) representa un avance considerable.

5.5 Discusión de Resultados

Los resultados experimentales presentados en este capítulo constituyen la evidencia empírica que permite evaluar el grado de cumplimiento de los objetivos de investigación. Esta sección analiza e interpreta los hallazgos principales a la luz de dichos objetivos, contextualiza los resultados en el marco de la literatura existente, y aborda el quinto objetivo específico (*OE5*): identificar fortalezas, limitaciones y áreas de mejora en el modelo y su aplicación en el mundo real

5.5.1 Interpretación de los Resultados Principales

El modelo *BSC-BIO-EHR-ES* mostró un desempeño sólido en la clasificación de textos médicos en español (*F1=0.7269, ROC-AUC=0.8058, PR-AUC=0.8565*), pese a tres condiciones relevantes: un *Dataset* pequeño de 534 muestras, un dominio altamente especializado como el cáncer tiroideo y el uso de recursos accesibles como *Google Colab Pro* con *GPU Tesla T4*. Esto sugiere una transferencia efectiva del conocimiento preentrenado y demuestra que es posible obtener resultados competitivos sin infraestructura costosa. Además, el balance entre *Precisión* (0.8725) y *Recall* (0.6519) refleja una configuración conservadora adecuada para apoyo clínico, al priorizar la reducción de falsos positivos. Aunque el *Recall* puede mejorar, el sistema resulta útil como herramienta complementaria al juicio médico. Este equilibrio también puede ajustarse modificando el umbral de clasificación según el contexto clínico, ya sea para priorizar *Recall* en tamizaje o precisión en confirmación diagnóstica.

5.5.2 Comparación con el Estado del Arte

La superioridad de *BSC-BIO-EHR-ES* frente a los algoritmos tradicionales de clasificación, con una mejora del 30–40% en *F1-Score*, respalda la hipótesis central del estudio y coincide con la tendencia reportada en la literatura sobre modelos de lenguaje aplicados a medicina. Aunque modelos biomédicos como *BioBERT* en inglés han alcanzado *F1-Scores* más altos, suelen entrenarse con *Datasets* mucho mayores; por ello, el resultado obtenido aquí (*F1=0.7269* con 534 muestras) sugiere una eficiencia de datos competitiva. Además, este trabajo constituye uno de los primeros esfuerzos en aplicar modelos de lenguaje biomédicos en español al diagnóstico de cáncer tiroideo, aportando una línea base para futuras investigaciones. La mejora de 30.6% sobre *Regresión Logística* y el hecho de que incluso *Random Forest* no se acercara al rendimiento de *BERT* evidencian una limitación clave de los enfoques basados en variables tabulares: no capturan toda la riqueza semántica de las narrativas clínicas. En cambio, *BERT* puede identificar patrones lingüísticos y contexto clínico difíciles de representar con características numéricas predefinidas.

5.5.3 Robustez y Estabilidad del Modelo

El análisis de estabilidad mostró una variabilidad moderada entre *folds* (*F1-Score: 0.6000–0.8085; σ=0.0622*), algo esperable dado el tamaño reducido del *Dataset* y la heterogeneidad propia de los casos clínicos. Aun así, el uso de *Nested Cross-Validation* con *10 folds* externos permitió obtener una estimación confiable del rendimiento promedio, y el modelo mantuvo un desempeño superior al azar en todos los *folds*. Además, el sesgo casi nulo entre validación interna y evaluación externa (*Δ=-0.0193*) indica que el procedimiento produjo estimaciones prácticamente no sesgadas, incluso ligeramente conservadoras. Por último, la baja variabilidad en *ROC-AUC* (*σ=0.0523*) refuerza que la capacidad discriminativa del modelo es estable y predecible, un aspecto especialmente valioso en aplicaciones clínicas.

5.5.4 Implicaciones para la Práctica Clínica

Los resultados sugieren que los modelos de lenguaje biomédico en español pueden ser una herramienta útil de apoyo al diagnóstico de cáncer tiroideo. Aunque su implementación práctica requiere consideraciones adicionales, el desempeño obtenido justifica explorar esta vía. El sistema permite ajustar el umbral de decisión según el contexto clínico, priorizando mayor *Recall* en screening o mayor precisión en confirmación diagnóstica. Además, técnicas de interpretabilidad pueden aportar mayor transparencia sobre los fragmentos del texto que influyen en cada predicción. En la práctica, el modelo podría integrarse en sistemas de historias clínicas electrónicas como apoyo para priorizar casos y, por su tamaño manejable, implementarse en servidores locales, favoreciendo el cumplimiento de requisitos de privacidad al evitar el envío de datos sensibles a servicios externos.

5.5.5 Significancia del Trabajo Realizado

Este trabajo aporta de forma relevante al campo de la inteligencia artificial médica en el contexto hispanohablante al demostrar que un modelo de lenguaje biomédico en español puede alcanzar un rendimiento competitivo (*F1=0.7269*) aun con un *Dataset* limitado de 534 muestras. Además, los análisis estadísticos aportan evidencia sólida de su superioridad frente a métodos tradicionales ($p < 10^{-13}$), mientras que la metodología basada en *Nested Cross-Validation* asegura estimaciones rigurosas y replicables. Otro aspecto destacable es que los resultados se obtuvieron con infraestructura accesible, lo que demuestra que este tipo de avances no depende necesariamente de recursos industriales. Finalmente, el trabajo tiene valor específico para el contexto mexicano al considerar idioma, terminología y regulación local, y establece una base metodológica y de rendimiento útil para futuras investigaciones.

En síntesis, los resultados experimentales demuestran el cumplimiento del objetivo general de esta investigación: se logró identificar características de la sintomatología del cáncer de tiroides mediante el reentrenamiento del modelo *BSC-BIO-EHR-ES*, generando información diagnóstica validada estadísticamente. Los objetivos específicos fueron cumplidos de manera secuencial y verificable. En primer lugar, el *OE1* se atendió mediante la compilación del corpus conforme a los criterios establecidos por la *NOM-004-SSA3-2012*, lo que permitió organizar la información bajo un marco normativo vigente. Posteriormente, el *OE2* permitió identificar características y limitaciones del *Dataset*, entre ellas su tamaño reducido, aspecto relevante porque puede influir en el desempeño y la generalización de los modelos. En relación con el *OE3*, la optimización del modelo se evidenció mediante los resultados obtenidos, alcanzando un *F1-score* de *0.7269* y un *ROC-AUC* de *0.8058*. Estos valores muestran un desempeño competitivo dentro de las condiciones del estudio. Para el *OE4*, los resultados del modelo basado en *LLM* fueron comparados con métodos de referencia, confirmando una mejora estadísticamente significativa frente a dichos enfoques, con un valor de $p\text{-value} < 10^{-13}$. Finalmente, el *OE5* permitió identificar las principales fortalezas y limitaciones del trabajo, así como establecer líneas de mejora para investigaciones futuras, especialmente en relación con la ampliación del *Dataset*, la validación externa y la evaluación del modelo en escenarios clínicos más diversos.

El trabajo sienta una base sólida y metodológicamente rigurosa para el desarrollo de sistemas de apoyo al diagnóstico basados en modelos de lenguaje especializados en el contexto médico hispanohablante, demostrando que resultados competitivos son alcanzables incluso con recursos limitados cuando se aplica una metodología rigurosa y se aprovechan adecuadamente los avances en modelos de lenguaje preentrenados.

CAPÍTULO VI: CONCLUSIONES

El presente capítulo cierra el trabajo de investigación articulando, en un solo cuerpo argumental, los hallazgos, el grado de cumplimiento de los objetivos propuestos, las aportaciones científicas y metodológicas derivadas del estudio, las limitaciones observadas y las líneas de trabajo futuro que se desprenden de los resultados. Con esta estructura se busca proporcionar al lector una síntesis conclusiva que permita valorar, en términos cuantitativos y cualitativos, el alcance real del reentrenamiento del modelo *BSC-BIO-EHR-ES* para la clasificación diagnóstica de cáncer de tiroides a partir de narrativas clínicas en español, así como identificar con precisión lo que ha quedado resuelto, lo que ha quedado acotado y lo que debe seguir explorándose.

6.1 Conclusiones generales

La investigación abordó la ausencia de herramientas de análisis automático de texto clínico en español para diagnóstico oncológico, una brecha asociada a tres limitantes: escasez de corpus clínicos abiertos en español, fragmentación deficiente del vocabulario médico por *tokenizadores* anglófonos (Carrino et al., 2022) y falta de modelos entrenados con datos mexicanos conforme a la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012). Para atenderla, se ejecutó un protocolo de cuatro fases: construcción de un corpus balanceado de 534 muestras, evaluación tabular con seis algoritmos clásicos, conversión tabular-textual verificada con *DeepSeek R1* en modo *SOTA* (Guo et al., 2025) y ajuste fino supervisado de *BSC-BIO-EHR-ES* (Carrino et al., 2022) mediante validación cruzada anidada de 10 pliegues externos y 3 internos.

Los resultados demostraron, bajo el mismo corpus y condiciones controladas, que la representación textual en lenguaje natural procesada por un modelo biomédico preentrenado en español aporta mayor valor diagnóstico que la representación tabular con clasificadores clásicos. El modelo ajustado obtuvo un *F1-Score* de 0.7269 ± 0.0622, *ROC-AUC* de 0.8058 ± 0.0523 y *PR-AUC* de 0.8565 ± 0.0339 en el bucle externo, superando tanto a la mejor línea base tabular Regresión Logística, $F1 = 0.5526 \pm 0.0742$ como al umbral tentativo $F1 \geq 0.70$ definido a priori (Mosbach et al., 2021).

Se obtuvieron tres hallazgos centrales. Primero, en este corpus la información diagnóstica no está en la topología vectorial de las 22 variables, sino en su articulación semántica: *K-Means* y *HDBSCAN* descartaron separabilidad natural y los seis clasificadores supervisados confirmaron que la representación tabular directa apenas supera el azar. Segundo, el ajuste fino completo de un modelo de 125 millones de parámetros sobre 534 muestras, con una ratio parámetro-muestra de $\sim 2\times10^{5}$:1, es viable si se combina con *Early Stopping* (*patience=2*), *Weight Decay 0.01*, *Dropout 0.1* y validación anidada estricta (Mosbach et al., 2021), (Demšar, 2006). Tercero, la superioridad del modelo no se explica por azar: la prueba de *Friedman* arrojó $\chi^2 = 186.81$ y $p = 1.21\times10^{-37}$, y las pruebas de *Wilcoxon* pareadas mostraron *p-values* $< 1.56\times10^{-13}$ en todas las comparaciones, incluso tras la corrección de *Bonferroni* (Demšar, 2006).

En síntesis, la tesis quedó verificada empíricamente dentro de los límites del corpus y del contexto experimental, y mostró que transformar datos tabulares en narrativas clínicas bajo *NOM-004* para activar conocimiento biomédico preentrenado constituye una estrategia válida y reproducible para diagnóstico asistido en escenarios de escasez de datos en español.

6.2 Objetivos alcanzados

La verificación del cumplimiento de los objetivos se realiza de forma explícita y trazable, vinculando cada objetivo específico con la fase metodológica en la que se operó y con el indicador empírico que sustenta su satisfacción. La Figura 6.1 sintetiza esta trazabilidad de manera visual, y la Tabla 6.1 formaliza el grado de cumplimiento alcanzado junto con la evidencia cuantitativa que lo respalda.

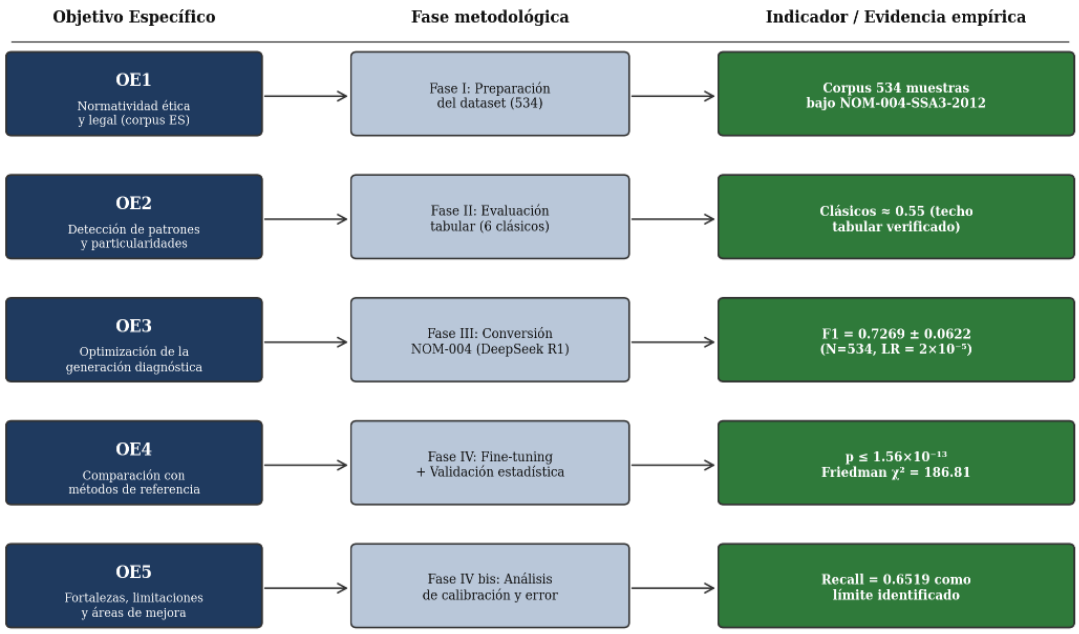

Figura 6.1: Mapa de trazabilidad entre objetivos específicos (OE1–OE5), fases metodológicas e indicadores empíricos que respaldan su cumplimiento.

El objetivo general se considera cumplido: se reentrenó y validó un modelo *BSC-BIO-EHR-ES* sobre un *corpus Gold Standard* de 534 narrativas clínicas, cuya capacidad diagnóstica fue caracterizada mediante métricas discriminativas (*F1, ROC-AUC, PR-AUC, MCC*), probabilísticas (*Brier Score*) y comparativas (*Wilcoxon*, *Friedman*) bajo validación anidada reproducible con semilla fija.

Los objetivos específicos también se cumplieron. El *OE1* se satisfizo con el uso exclusivo de datos abiertos del programa *SEER* del *National Institutes of Health* (National Institutes of Health, "Surveillance, Epidemiology, and End Results (SEER) Program", 2023), la adecuación de las narrativas a la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012) y la declaración explícita de *DeepSeek R1* como herramienta de conversión textual conforme a los principios de transparencia de *Ellaway* y *Tolsgaard* (Ellaway & Tolsgaard, 2023). El *OE2* se verificó al demostrar que el *Dataset* de 534 muestras está por debajo del umbral de 2,491 ejemplos reportado por Mosbach (Mosbach et al., 2021) para la estabilidad del fine-tuning de *BERT*, y que el espacio tabular carece de separabilidad natural, ya que ningún algoritmo clásico superó significativamente el azar, justificando la transición a representación semántica. El *OE3* se materializó en los resultados del modelo: *F1 = 0.7269*, *ROC-AUC = 0.8058*, *PR-AUC = 0.8565*, *MCC = 0.5345*, *Precisión = 0.8725* y *Recall = 0.6519*, superiores al umbral $F1 \geq 0.70$ y consistentes con una transferencia efectiva del conocimiento biomédico preentrenado a partir del corpus original de 95 millones de *tokens* de *BSC-BIO-EHR-ES* (Carrino et al., 2022). El *OE4* se cumplió al comparar el modelo con seis algoritmos clásicos *Regresión Logística*, *Random Forest*, *SVM-RBF*, *Árbol de Decisión*, *KNN y Naive Bayes* bajo el mismo esquema de validación anidada y con pruebas de *Friedman* y *Wilcoxon* (Demšar, 2006), que confirmaron diferencias estadísticamente significativas y prácticamente relevantes. El *OE5* se satisfizo mediante el análisis de estabilidad *inter-fold* ($\sigma = 0.0622$), del sesgo entre validación interna y externa ($\Delta = -0.0193$, indicativo de una estimación conservadora) y de la tasa de falsos negativos (34.8%), identificando al *Recall* como principal área de mejora.

Las hipótesis también quedaron verificadas. La *HG* se sostuvo con una ventaja mínima de $+0.1691$ en frente al mejor clasificador tabular, estadísticamente significativa con $p \leq 1.56 \times 10^{-13}$ y de tamaño de efecto grande según Cohen. La *H1* se confirmó con un *F1* promedio de *0.7269* y una desviación estándar de *0.0622*,

ambos dentro de los umbrales establecidos. La *H2* se verificó porque una tasa de aprendizaje de 2×10^{-5} produjo convergencia estable, evidenciada por curvas paralelas de *train_loss* y *eval_loss* sin divergencia pronunciada y por una selección consistente de la época óptima mediante *Early Stopping*. La Tabla 6.2 resume esta verificación.

Tabla 6.1. Verificación del cumplimiento de los objetivos específicos con indicador y evidencia empírica.

Objetivo	Indicador asociado	Resultado obtenido	Grado de cumplimiento
OE1	Observar la normatividad ética y legal vigente para compilar y gestionar corpus de textos en español relacionados con los casos de cáncer en México, incluyendo diagnóstico, tratamiento y resultados clínicos.	Corpus de 534 narrativas *NOM-004* con doble filtro (*DeepSeek R1 + Agente Supervisor*) y datos *SEER* de acceso abierto	Cumplido
OE2	Identificación de patrones y limitantes que afectan la generación	Separabilidad tabular descartada (*K-Means*/*HDBSCAN*); 6 clasificadores clásicos ≤ 0.55 *F1*; $N=534$ < umbral de 2,491 (Mosbach et al., 2021)	Cumplido
OE3	Optimizar la capacidad de generar información sobre el diagnóstico del cáncer en español	*F1 = 0.7269*, *ROC-AUC = 0.8058*, *PR-AUC = 0.8565*, *MCC = 0.5345* (*Nested CV 10×3*)	Cumplido ($F1 \geq 0.70$)
OE4	Comparar formal con métodos de referencia	*Wilcoxon* $p \leq 1.56\times10^{-13}$ en las 6 comparaciones; *Friedman* $\chi^2 = 186.81$, $p = 1.21\times10^{-37}$	Cumplido
OE5	Caracterización de fortalezas, limitaciones y áreas de mejora	*Precisión 0.8725* (fortaleza); *Recall 0.6519* y *FN = 34.8%* (limitación); $\sigma F1 = 0.0622$ (estabilidad)	Cumplido

Tabla 6.2. Verificación empírica de las hipótesis de trabajo formuladas a priori.

Hipótesis	Postulado	Resultado empírico	Veredicto
HG	El modelo ajustado obtendrá *F1* superior al de clasificadores tabulares sobre la misma información	*F1 = 0.7269* vs máximo tabular *0.5526* (*Reg. Logística*). $\Delta = +0.1693$. *Wilcoxon* $p = 1.56\times10^{-13}$	Verificada
H1	*Nested CV (10×3) + Early Stopping* mantendrán varianza controlada con $F1 \geq 0.7$	*F1* promedio 0.7269; $\sigma F1 = 0.0622 \leq 0.10$; *Inner-Outer bias* = –0.0193	Verificada
H2	$LR = 2\times10^{-5}$ permitirá convergencia estable sin divergencia	*Curvas Train/Val* paralelas; *Early Stopping* consistente entre *folds*; sin patrón *X-shape* (observado en la Figura 4.5)	Verificada

6.3 Aportaciones

La investigación produjo aportaciones en tres dimensiones diferenciadas, cuya articulación conjunta constituye el valor diferencial del trabajo respecto al estado del arte revisado en el Capítulo 3.

6.3.1 Aportación metodológica

La primera aportación, la más transferible a otros dominios clínicos, es un protocolo verificable de conversión tabular-textual que resuelve tres problemas detectados en la fase exploratoria: la degradación cognitiva de modelos locales cuantizados, el desajuste entre preentrenamiento narrativo y tarea decodificadora, y la dependencia circular cuando un mismo modelo genera y verifica el texto. El pipeline combina inferencia por *API* con *DeepSeek R1* en modo *SOTA* (Guo et al., 2025), *Cadena de Pensamiento* (*CoT*) auditable mediante la *Traza de Razonamiento*, y un *Agente Supervisor* asimétrico al que se inyecta el *Ground Truth* tabular con etiquetas *XML* bajo una regla de prioridad ontológica que privilegia los marcadores sobre la narrativa. Al separar explícitamente generación y verificación, esta arquitectura puede extenderse a cualquier dominio médico que requiera convertir datos estructurados en narrativas ajustadas a una norma de redacción específica, como la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012).

La segunda aportación metodológica es un esquema robusto de validación estadística para clasificación binaria con datos limitados. La combinación de validación cruzada anidada de 10 pliegues externos y 3 internos

con las pruebas de *Friedman* y *Wilcoxon*, más la corrección de Bonferroni para comparaciones múltiples, produjo niveles de certeza con *p-values* del orden de 10^{-15}, superiores a los estándares habituales de la literatura revisada, dominada por *hold-out simple* o *k-fold* no anidado. Además, el esquema incorpora un criterio explícito de sesgo interno-externo ($\Delta < 2\text{-}3\%$) como indicador de integridad de la estimación, útil para distinguir entre rendimiento real y sesgo optimista de selección.

6.3.2 Aportación técnica

Desde el plano técnico, la aportación central es la evidencia de que el ajuste fino completo de un *Transformer* de 125 millones de parámetros sobre solo 534 muestras es viable en un dominio altamente especializado si se aplica el régimen de regularización descrito: *Early Stopping* con *patience=2*, *Weight Decay 0.01*, Dropout 0.1 y $LR = 2\times10^{-5}$ con *AdamW* (Loshchilov & Hutter, 2019). En el contexto del español biomédico, este resultado contradice la frontera de estabilidad reportada por Mosbach (Mosbach et al., 2021) para inglés general y sugiere que un preentrenamiento en datos de dominio muy cercano a la tarea final, junto con hiperparámetros conservadores, puede compensar parcialmente la escasez de datos. La diferencia *Inner-Outer* de *–0.0193* no solo valida el esquema de evaluación, sino que indica estimaciones prácticamente no sesgadas e incluso ligeramente conservadoras.

La segunda aportación técnica es que se obtuvo un resultado estadísticamente competitivo sin infraestructura de cómputo industrial. Todo el *pipeline* conversión tabular-textual, fine-tuning completo y validación estadística se ejecutó en *Google Colab Pro* con una sola *GPU Tesla T4*, lo que favorece la reproducibilidad y su replicación en contextos institucionales con recursos limitados, incluidos hospitales públicos y centros de investigación del sistema de salud mexicano.

6.3.3 Aportación contextual

En el plano contextual, el trabajo constituye, hasta donde alcanza la revisión del Capítulo 3, uno de los primeros esfuerzos documentados en aplicar un modelo de lenguaje biomédico nativo en español al diagnóstico específico de cáncer de tiroides. Aborda simultáneamente tres brechas del estado del arte: la lingüística, al operar íntegramente en español biomédico con un modelo preentrenado en 95 millones de *tokens* de *EHR* reales en ese idioma (Carrino et al., 2022); la de dominio oncológico, al extender su uso más allá del reconocimiento de entidades nombradas hacia la clasificación diagnóstica binaria; y la de validación estadística, al incorporar pruebas no paramétricas formales con corrección por comparaciones múltiples. Además, el ajuste del pipeline de generación textual a la *NOM-004-SSA3-2012* (Secretaría de Salud, 2012)ancla el trabajo al contexto normativo mexicano y lo hace más transferible a la práctica clínica local que adaptaciones generalistas en español peninsular o multilingüe. Como aportación exploratoria adicional, el trabajo deja formulada y parcialmente implementada una arquitectura híbrida de aumento de datos en espacio latente, donde los *embeddings* contextuales de *BSC-BIO-EHR-ES* sirven de sustrato para un Algoritmo Genético Difuso (*DEAP por sus siglas en ingles*) cuya función de aptitud se basa en la Ecuación 6.1, adaptada de la formulación clásica de conjuntos difusos de Zadeh (Zadeh, 1965):

$$\mu\_c(x) = exp(\frac{-\|x - \mu\_c\|^2}{2\sigma\_c^2}) \quad (6.1)$$

donde μ_c es el centroide de los vectores reales de la clase c y σ_c la dispersión media; ambos permiten penalizar exponencialmente los vectores sintéticos alejados del clúster clínico real. Aunque la comparación final entre la variante base y la optimizada con aumento sintético no es el eje central del trabajo, la arquitectura documentada abre una línea replicable para escenarios de escasez de datos aún más severa.

6.4 Limitaciones

La identificación explícita de las limitaciones del trabajo es un requisito metodológico indispensable, no solo para calibrar correctamente el alcance de las conclusiones alcanzadas, sino para delimitar con precisión el terreno sobre el que deberá operar la investigación subsecuente. Las limitaciones observadas se agrupan en cuatro categorías interrelacionadas.

6.4.1 Limitación por tamaño del corpus

La restricción más determinante del estudio es el tamaño del corpus efectivo. El *Dataset SEER* original contenía únicamente 267 casos positivos confirmados de cáncer de tiroides, lo cual fijó, por construcción, el techo de la clase minoritaria y obligó a trabajar con una muestra analítica de 534 registros tras el submuestreo balanceado. Esta condición impone un techo empírico a la capacidad de generalización del modelo independientemente de la sofisticación del ajuste fino, y explica en buena medida la variabilidad *inter-fold* observada (*F1* oscilando entre *0.6* y *0.8085* a lo largo de los 10 pliegues externos). Aun cuando la varianza global ($\sigma F1 = 0.0622$) permaneció dentro del umbral metodológico aceptable, el rango de rendimiento por pliegue evidencia que la composición particular de cada partición todavía influye de forma no despreciable en los resultados.

6.4.2 Limitación por desequilibrio entre precisión y recall

El modelo ajustado muestra un balance asimétrico entre precisión (0.8725) y recall (0.6519), lo que implica una tasa de falsos negativos del 34.8%: con el umbral por defecto, cerca de uno de cada tres casos malignos no sería detectado. En un contexto clínico real, esto limita su uso directo como herramienta de tamizaje y lo orienta más hacia la confirmación diagnóstica o la priorización de casos dentro de flujos clínicos existentes. Aunque rebalancear el umbral de decisión es técnicamente simple, ello desplaza el problema hacia los falsos positivos, y optimizar ambas dimensiones exige más datos que los disponibles en este estudio.

6.4.3 Limitación por ausencia de validación externa y generalización

El modelo se evaluó únicamente sobre particiones del mismo corpus SEER usado en el entrenamiento, mediante validación cruzada anidada. Aunque este esquema ofrece estimaciones no sesgadas dentro del corpus, su generalización a otras poblaciones con distintas distribuciones demográficas, estilos de redacción clínica, sistemas de salud o terminología local no puede asumirse sin validación multicéntrica posterior. Además, el entrenamiento se realizó sobre narrativas sintetizadas a partir de datos tabulares y no sobre historias clínicas primarias redactadas por profesionales de la salud; por ello, el pipeline de conversión introduce un sesgo estilístico que, aunque atenuado por la verificación del Agente Supervisor, no reproduce la variabilidad expresiva de un corpus clínico primario.

6.4.4 Limitación por alcance clínico del artefacto

El sistema se plantea explícitamente como investigación aplicada de desarrollo tecnológico de base, no como una solución de despliegue clínico inmediato. No ha sido validado prospectivamente con oncólogos, carece de aprobación regulatoria y no incorpora módulos de interpretabilidad por *token*, auditoría de decisiones ni mecanismos de retracción, todos necesarios para su uso asistencial. Además, el estudio se limita al diagnóstico binario de cáncer de tiroides y no aborda la subtipificación oncológica papilar, folicular, medular, anaplásico, clínicamente relevante para la decisión terapéutica. La extensión multiclase queda, por tanto, como una línea natural de trabajo futuro.

6.5 Trabajo futuro

Las limitaciones identificadas y los hallazgos alcanzados perfilan una agenda de trabajo futuro estructurada en cuatro ejes que progresan desde lo técnicamente inmediato hasta lo institucionalmente más ambicioso.

6.5.1 Expansión y consolidación del corpus clínico

La prioridad inmediata es ampliar el corpus con historias clínicas primarias en español de instituciones mexicanas, bajo convenios que aseguren cumplimiento de la *NOM-004-SSA3-2012*(Secretaría de Salud, 2012), anonimización y consentimiento informado. Un corpus de varios miles de muestras permitiría elevar el techo empírico del estudio y evaluar si la tendencia observada alta *Precisión* y *Recall* comprometido se mantiene o se corrige con más datos. En paralelo, debe incorporarse redacción clínica auténtica, no mediada por el pipeline de conversión, para reducir el sesgo estilístico señalado en la sección 6.4.3.

6.5.2 Extensión multiclase y tareas clínicamente más ricas

La segunda línea consiste en transitar del diagnóstico binario hacia la subtipificación histopatológica del cáncer tiroideo (papilar, folicular, medular, anaplásico), con su correspondiente reorganización del corpus y de los indicadores de evaluación. De manera complementaria, debe explorarse la extensión del modelo a tareas de extracción de entidades clínicas (fechas, dosis, procedimientos, antecedentes), pronóstico de evolución y estratificación de riesgo, todas ellas pertinentes al flujo de trabajo oncológico real.

6.5.3 Validación clínica prospectiva y auditoría de calibración

La tercera línea es la validación clínica prospectiva con oncólogos tiroideos del sistema de salud mexicano, para contrastar las predicciones del modelo con el juicio experto en casos inéditos y evaluar su calibración probabilística en condiciones reales. En este contexto, la honestidad probabilística medida mediante *Brier Score* y *ECE* (C. Guo et al., 2017) es crítica, ya que los médicos interpretarán las probabilidades emitidas como grados de certeza. Esta línea exige además mecanismos de interpretabilidad a nivel de *token* atribución de características y pruebas de estrés lingüístico que permitan explicar qué fragmentos del texto clínico sustentan cada predicción.

6.5.4 Integración institucional y productos académicos derivados

Como cierre institucional, se prevé publicar el modelo ajustado y el pipeline de conversión bajo licencia abierta en el repositorio público del grupo, en coherencia con el criterio de reproducibilidad que guio el estudio. La memoria técnica, las figuras experimentales, los scripts de evaluación estadística y los *tests* de estrés lingüístico servirán de base para al menos un artículo en congreso nacional y, eventualmente, una publicación en una revista de informática médica en español. Además, el protocolo de endurecimiento del servidor autenticación por llave asimétrica, migración de *SSH* al puerto *5022*, *firewall UFW* de mínimo privilegio e intercambio *Curve25519* con cifrado *ChaCha20-Poly1305*, activado tras el incidente de seguridad detectado en la fase exploratoria, queda documentado como contribución operativa para futuros proyectos con datos médicos sensibles. La integración del modelo en un sistema conversacional multivuelta con razonamiento estructurado (*CoT, ReAct*) y guías clínicas externas, siguiendo las arquitecturas de Abbasian et al. y Wang et al. revisadas en el Capítulo 3, perfila su evolución hacia un asistente diagnóstico interactivo desplegable en servidores locales y compatible con los requisitos de privacidad y soberanía de datos del sistema de salud mexicano.

En síntesis, el trabajo cierra una etapa del programa de investigación demostrar empírica y estadísticamente que un modelo de lenguaje biomédico nativo en español puede superar significativamente a los métodos clásicos en la clasificación diagnóstica de cáncer tiroideo, incluso con severa escasez de datos y abre una ruta ordenada hacia herramientas de apoyo a la decisión clínica con validez metodológica, normativa y contextual para el sistema de salud hispanohablante.

REFERENCIAS BIBLIOGRÁFICAS

- Abbasian, M., Azimi, I., Rahmani, A. M., & Jain, R. (2024). *Conversational Health Agents: A Personalized LLM-Powered Agent Framework*. https://doi.org/10.48550/arXiv.2310.02374
- Agerri, R., Borber, I., Campos, J., Gonzalez-Agirre, A., Gurrutxaga, A., Soroa, A., & Ferrández, A. (2023). Challenges and solutions in multilingual clinical NLP". *En Proc. Clinical Natural Language Processing Workshop*, 1-12.
- Alghamdi, H. M., & Mostafa, A. (2024). Towards Reliable Healthcare LLM Agents: A Case Study for Pilgrims during Hajj. *Information*, *15*(7), 371. https://doi.org/10.3390/info15070371
- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings". *En Proc. 2nd Clinical Natural Language Processing Workshop*, 72-78.
- Asociación Española Contra el Cáncer. (s. f.). *Cáncer de tiroides: Diagnóstico*. Recuperado https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-tiroides/diagnostico
- Banderas, E. R. (2017). Detección del anillo de colesterol en el iris mediante procesamiento digital de imágenes en dispositivos móviles". En *Tesis de maestría, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET*.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed., Número BN). Springer.
- Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. (2024). Machine learning for risk stratification of thyroid cancer patients: A 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2095-2104. https://doi.org/10.1007/s00405-023-08299-w
- Borzooei, S., & Tarokhian, A. (2023). *Differentiated Thyroid Cancer Recurrence [Conjunto de datos]*. https://doi.org/10.24432/C5632J
- Brown, T., Mann, B., Ryder, N., Subbiah, M., & Kaplan, J. D. (2020). Language models are few-shot learners". En *En Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877-1901).
- Carrino, C. P., Armengol-Estapé, J., Gutiérrez-Fandiño, A., Llop-Palao, J., Pàmies, M., Gonzalez-Agirre, A., & Villegas, M. (2021). *Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario* (Versión 2). arXiv. https://doi.org/10.48550/ARXIV.2109.03570
- Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Valencia, A., Gonzalez-Agirre, A., & Villegas, M. (2022). Pretrained biomedical language models for clinical NLP in Spanish". *En Proc. 21st Workshop Biomedical Language Processing, Dublín, Irlanda*, 193-199.
- Caruana, R., Lawrence, S., & Giles, C. L. (2000). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping". En *Advances in Neural Information Processing Systems* (Vol. 13).
- 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,.
- Chen, Y., Wang, Z., Xing, X., Zheng, H., Xu, Z., Fang, K., Li, S., Wang, J., Wu, J., Liu, Q., & Xu, X. (2023). *BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT*. https://doi.org/10.48550/arXiv.2310.15896
- Collins, G. S., Moons, K. G. M., Dhiman, P., & others. (2024). TRIPOD+AI statement. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378
- Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*(1), 1-30,.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En *Proc* (pp. 4171-4186,). NAACL-HLT. https://doi.org/10.18653/v1/N19-1423.
- Ellaway, R. H., & Tolsgaard, M. (2023). Artificial scholarship: LLMs in health professions education research. *Advances in Health Sciences Education*, *28*(3), 659-664. https://doi.org/10.1007/s10459-023-10257-4
- Ellaway, R., & Tolsgaard, M. (2023). Artificial scholarship: The implications of large language models for health professions education research". En *Advances in Health Sciences Education* (Vol. 28, pp. 659-664,).
- Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement". *ACM SIGKDD Explorations Newsletter*, *12*(1), 49-57,.
- Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019). PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track". *En Proc. 5th Workshop BioNLP Open Shared Tasks*, 1-10.
- Goyal, S., Rastogi, E., Rajagopal, S. P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., & Ward, J. (2024). HealAI: A Healthcare LLM for Effective Medical Documentation. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, 1167-1168. https://doi.org/10.1145/3616855.3635739
- Gretton, A., Borgwardt, K., Rasch, M. J., Scholkopf, B., & Smola, A. J. (2008). *A Kernel Method for the Two-Sample Problem* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.0805.2368

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks". *en Proc. ICML*, 1321-1330.
- Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*(8081), 633-638. https://doi.org/10.1038/s41586-025-09422-z
- Gururajan, A. K., Lopez-Cuena, E., Bayarri-Planas, J., Tormos, A., Hinjos, D., Bernabeu-Perez, P., Arias-Duart, A., Martin-Torres, P. A., Urcelay-Ganzabal, L., Gonzalez-Mallo, M., Alvarez-Napagao, S., Ayguadé-Parra, E., Cortés, U., & Garcia-Gasulla, D. (2024). *Aloe: A Family of Fine-tuned Open Healthcare LLMs*. https://doi.org/10.48550/arXiv.2405.01886
- Hao, Y., Dong, L., Wei, F., & Xu, K. (2019). Visualizing and Understanding the Effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019)*. https://arxiv.org/abs/1908.05620
- He, Y., Zhu, Z., Zhang, Y., Chen, Q., & Caverlee, J. (2020). Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4604-4614. https://doi.org/10.18653/v1/2020.emnlp-main.372
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification". *En Proc. ACL*, 328-339.
- Hundahl, S. A., Fleming, I. D., Fremgen, A. M., & Menck, H. R. (1998). A National Cancer Data Base report on 53,856 cases of thyroid carcinoma treated in the U.S., 1985–1995". *Cancer*, *83*(12), 2638-2648,.
- I.N.E.G.I. (2024). Estadísticas a propósito del Día Mundial contra el Cáncer (4 de febrero)". En *Instituto Nacional de Estadística y Geografía, Comunicado de prensa núm. 143/24, México*. https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2024/EAP_CANCER24.pdf
- J.A.Pineda Tapia. (2016). Desarrollo de una aplicación bajo la plataforma Android para la detección de cáncer abdominal mediante el análisis de alteraciones en el iris". En *Tesis de maestría, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET*.
- Jensen, P., Jensen, L., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care". *Nature Reviews Genetics*, *13*(6), 395-405,.
- Joseph-Luna, J., Rodríguez-Palomares, L. A., Olvera-Juárez, M., Reynoso-Noverón, N., & Pacheco-Bravo, I. (2014). Validez y precisión del ultrasonido como método diagnóstico del cáncer de tiroides en pacientes del Instituto Nacional de Cancerología, México. *Gaceta Mexicana de Oncología*, *13*(6), 388-396. https://doi.org/10.1016/j.gamo.2014.06.003
- Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing* (3rd ed.).
- Kim, Y., Xu, X., McDuff, D., Breazeal, C., & Park, H. W. (2024). Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. *Proceedings of the 5th Conference on Health, Inference, and Learning (CHIL), Proceedings of Machine Learning Research*, *248*, 522-539.
- Kirk, S., Lee, Y., Roche, C., Bonaccio, E., Filippini, J., & Jarosz, R. (2016). *The Cancer Genome Atlas Thyroid Cancer Collection (TCGA-THCA)* (Versión 3) [Dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.9ZFRVF1B
- Koroteev, M. V. (2021). *BERT: A Review of Applications in Natural Language Processing and Understanding*. https://doi.org/10.48550/arXiv.2103.11943
- Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Jones Bell, M. (2024). The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, *11*, e59479. https://doi.org/10.2196/59479
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics*, *36*(4), 1234-1240,.
- Lee, R. Y., Kross, E. K., Torrence, J., Li, K. S., Sibley, J., Cohen, T., Lober, W. B., Engelberg, R. A., & Curtis, J. R. (2023). Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Network Open*, *6*(3), e231204. https://doi.org/10.1001/jamanetworkopen.2023.1204
- Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., & Yu, H. (2019). Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Medical Informatics*, *7*(3), e14830. https://doi.org/10.2196/14830
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.1907.11692
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization". *En Proc. ICLR*.

- Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579-2605,.
- Meskó, B. (2023). The Impact of Multimodal Large Language Models on Health Care's Future. *Journal of Medical Internet Research*, *25*, e52865. https://doi.org/10.2196/52865
- Miranda-Escalada, A., Farr'e-Maduell, E., & Krallinger, M. (2020). Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish. *Proceedings of IberLEF 2020, CEUR-WS Vol-2664*, 303-323.
- Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Estrada, D., Gascó, L., & Krallinger, M. (2022). Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of LivingNER shared task and resources". *Procesamiento Del Lenguaje Natural*, *69*, 241-253,.
- Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., … Van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, *388*, e082505. https://doi.org/10.1136/bmj-2024-082505
- Morales Lozada, L. D., & Acosta Gavilánez, R. I. (2023). Prevención del cáncer de tiroides en atención primaria de salud. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, *4*(4), 1035-1051. https://doi.org/10.56712/latam.v4i4.1282
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines". *En Proc. ICLR*.
- *Learning Representations (ICLR 2021)*. https://arxiv.org/abs/2006.04884
- Mumtaz, U., Ahmed, A., & Mumtaz, S. (2023). *LLMs-Healthcare: Current Applications and Challenges of Large Language Models in various Medical Specialties*. https://doi.org/10.48550/arXiv.2311.12882
- *National Institutes of Health, "Surveillance, Epidemiology, and End Results (SEER) Program"*. (2023). National Cancer Institute. https://seer.cancer.gov
- Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges". *Journal of Biomedical Semantics*, *9*, 12,.
- Nguyen, K.-A.-N., Patel, D., Edalati, M., Sevillano, M., Timsina, P., Freeman, R., Levin, M. A., Reich, D. L., & Kia, A. (2025). Electronic-Medical-Record-Driven Machine Learning Predictive Model for Hospital-Acquired Pressure Injuries: Development and External Validation. *Journal of Clinical Medicine*, *14*(4), 1175. https://doi.org/10.3390/jcm14041175
- Niu, H., Omitaomu, O. A., Langston, M. A., Olama, M., Ozmen, O., Klasky, H. B., Laurio, A., Ward, M., & Nebeker, J. (2024). EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. *Journal of Biomedical Informatics*, *150*, 104605. https://doi.org/10.1016/j.jbi.2024.104605
- Organización Panamericana de la Salud. (2024). *Cáncer*. OPS/OMS. https://www.paho.org/es/temas/cancer
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). *Scikit-learn: Machine Learning in Python*. https://doi.org/10.48550/ARXIV.1201.0490
- Quinlan, R. (1986). *Thyroid Disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5D010
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, *4*(1), 86. https://doi.org/10.1038/s41746-021-00455-y
- Secretaría de Salud. (2012). *Norma Oficial Mexicana NOM-004-SSA3-2012, Del expediente clínico*. Diario Oficial de la Federación. https://dof.gob.mx/nota_detalle_popup.php?codigo=5272787
- Skianis, K., Pavlopoulos, J., & Doğruöz, A. S. (2024). *Building Multilingual Datasets for Predicting Mental Health Severity through LLMs: Prospects and Challenges* (Versión 3). arXiv. https://doi.org/10.48550/ARXIV.2409.17397
- Smith, A. R. (1978). Color gamut transform pairs". *ACM SIGGRAPH Computer Graphics*, *12*(3), 12-19,.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP". *En Proc. 57th Annu. Meeting Assoc. Comput*, 3645-3650.
- Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need". En *En Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008).
- Wang, Z., Dai, Z., P'oczos, B., & Carbonell, J. (2019). Characterizing and Avoiding Negative Transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. https://arxiv.org/abs/1811.09751

- Wang, Z., Li, H., Huang, D., Kim, H.-S., Shin, C.-W., & Rahmani, A. M. (2025). HealthQ: Unveiling questioning capabilities of LLM chains in healthcare conversations. *Smart Health*, 100570. https://doi.org/10.1016/j.smhl.2025.100570
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
- Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1-32. https://doi.org/10.1145/3643540
- Yu, H., Fan, L., Li, L., Zhou, J., Ma, Z., Xian, L., Hua, W., He, S., Jin, M., Zhang, Y., Gandhi, A., & Ma, X. (2024). Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis. *Journal of Healthcare Informatics Research*, *8*, 658-711. https://doi.org/10.1007/s41666-024-00171-8
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338-353,. https://doi.org/10.1016/S0019-9958(65)90241-X.
- Zhang, Q., Sun, Y., Zhang, L., Jiao, Y., & Tian, Y. (2021). Named entity recognition method in health preserving field based on BERT. *Procedia Computer Science, 183*, 212-220. https://doi.org/10.1016/j.procs.2021.02.052

ANEXOS

ANEXO A

Sección 1

La Tabla 7.1 del presente anexo presenta el diccionario completo de las 22 características predictoras, incluyendo su tipo de dato, codificación y descripción. Esta información es esencial para comprender la naturaleza de las variables que serán procesadas por los distintos métodos de clasificación.

Tabla 7.1: Diccionario de variables del dataset tabular (22 características predictoras).

#	Variable	Tipo	Descripción / Codificación
1	*age*	Numérica	Edad al momento de aleatorización
2	*sex*	Categórica binaria	Sexo (1=Masculino; 2=Femenino)
3	*race7*	Categórica	Raza/Etnia (1=Blanco no Hisp.; 2=Negro no Hisp.; 3=Hispano; 4=Asiático; 5=Isleño Pacífico; 6=Indio Amer.; 7=Desconocido)
4	*hispanic_f*	Categórica	Origen hispano (.F=Sin Formulario; .M=No Respondido; 0=No; 1=Sí)
5	*educat*	Categórica ordinal	Escolaridad (1=<8 años; 2=8-11; 3=Preparatoria; 4=Post-prep; 5=Univ. trunca; 6=Universitario; 7=Posgrado)
6	*occupat*	Categórica	Ocupación (1=Ama de Casa; 2=Trabajando; 3=Desempleado; 4=Retirado; 5=Baja por Enfermedad; 6=Discapacitado; 7=Otro)
7	*cig_stat*	Categórica	Estatus de tabaquismo (0=Nunca; 1=Fumador actual; 2=Exfumador)
8	*cig_years*	Numérica	Años fumando (decimal: .5 = 6 meses adicionales)
9	*pack_years*	Numérica	Paquetes-año de exposición al tabaco
10	*cigpd_f*	Categórica ordinal	Cigarrillos/día (0=0; 1=1-10; 2=11-20; 3=21-30; 4=31-40; 5=41-60; 6=61-80; 7=81+)
11	*thyd_fh*	Categórica	Antecedente familiar (1er grado) de cáncer de tiroides (0=No; 1=Sí; 9=Posible)
12	*thyd_fh_age*	Numérica	Edad del familiar más joven con cáncer de tiroides
13	*thyd_fh_cnt*	Numérica	Número de familiares (1er grado) con cáncer de tiroides
14	*fh_cancer*	Categórica binaria	Antecedente familiar de cualquier cáncer (0=No; 1=Sí)
15	*bmi_curc*	Categórica ordinal	*IMC* actual categorizado (1=<18.5; 2=18.5-25; 3=25-30; 4=>30)
16	*height_f*	Numérica	Altura en pulgadas
17	*weight_f*	Numérica	Peso en libras
18	*bmi_20c*	Categórica ordinal	*IMC* a los 20 años categorizado (misma codificación que *bmi_curc*)
19	*bmi_50c*	Categórica ordinal	*IMC* a los 50 años categorizado (misma codificación que *bmi_curc*)
20	*entryage_bq*	Numérica	Edad de entrada al estudio
21	*ph_thyd_bq*	Categórica	Antecedente personal de cáncer de tiroides (0=No; 1=Sí; 9=Desconocido)
22	*ph_any_bq*	Categórica	Antecedente personal de cualquier cáncer (0=No; 1=Sí; 9=Desconocido)

Sección 2

A. Variable Independiente (Factor de Tratamiento)

La variable independiente se define como la combinación de representación de datos y método de clasificación.

Implicación Metodológica: Esta estructura permite explorar dos preguntas de investigación: (1) ¿La representación textual podría aportar valor predictivo sobre la tabular? (2) ¿El fine-tuning específico podría superar a la "inteligencia general" de los *LLMs* en configuración *zero-shot*?

B. Variables Dependientes (Métricas de Desempeño)

Las variables de respuesta, presentadas en la Tabla 7.2, fueron seleccionadas para capturar múltiples dimensiones del rendimiento diagnóstico, evitando la dependencia de una única métrica.

Tabla 7.2: Variables dependientes (métricas de evaluación).

Métrica	Definición	Relevancia
F1-Score	Media armónica de *Precisión y Recall: 2·(P·R)/(P+R)*	Métrica principal. Balance entre *FP y FN*.
Exactitud	Proporción de predicciones correctas: *(TP+TN)/N*	Rendimiento global en *Dataset* balanceado.
Precisión	Verdaderos positivos entre predicciones positivas: *TP/(TP+FP)*	Minimiza alarmas falsas.
Sensibilidad	Positivos detectados entre reales: *TP/(TP+FN)*	Evita omisión de diagnósticos.
ROC-AUC	Área bajo la curva *ROC*	Capacidad discriminativa independiente del umbral.

C. Variables de Control

Para garantizar la validez interna del experimento y la comparabilidad entre métodos, se fijaron las variables de control documentadas en la Tabla 7.3.

Tabla 7.3: Variables de control del experimento.

Variable	Valor Fijado	Propósito
Tamaño del Dataset	534 registros (fijo)	Comparabilidad directa
Balance de Clases	50/50 (267/267)	Evitar sesgo de prevalencia
Semilla Aleatoria	*SEED = 42*	Reproducibilidad determinista
Validación Cruzada	*Stratified K-Fold (K=10)*	Preservar proporción en cada *fold*
Etiquetas	Idénticas para ambas representaciones	Correspondencia *1:1* tabular↔texto

Sección 3

En la Tabla 7.4 se muestra un condensado tanto de las hipótesis como de los criterios esperados.

Tabla 7.4: Resumen de hipótesis de trabajo y criterios de evaluación.

ID	Postulado	Criterio Tentativo
HG	Fine-tuning podría superar a clasificación tabular e igualar usando *few-shot* en modelos estado *SOTA*	$F1 \geq 0.70$ (estimado)
H1	Estabilidad razonable del rendimiento entre particiones	$\sigma\_F1 \leq 0.05$-0.10 (esperado)
H2	Convergencia estable sin divergencia de gradiente	Curvas acopladas, épocas 3-6

Sección 4-A

A. Infraestructura de Hardware

Los experimentos de clasificación tabular, fine-tuning y evaluación se ejecutaron en el entorno descrito en la Tabla 7.5.

Tabla 7.5: Especificaciones del hardware.

Componente	Especificación
Equipo	*COLAB Pro* (en línea)
Procesador	*Intel Xeon (2.0 GHz)*
Memoria RAM	*51 GB*
GPU	*NVIDIA Tesla T4*
Sistema Operativo	*Ubuntu 22.04.4 LTS*

B. *Stack* de Software

El *stack* de software utilizado se detalla en la Tabla 7.6, incluyendo las bibliotecas principales para clasificación, *deep learning* y análisis estadístico.

Tabla 7.6 Anexo A: Stack de software.

Componente	Versión	Función
Python	3.10+	Lenguaje principal
Scikit-learn	1.3+	Algoritmos de clasificación, preprocesamiento, métricas
Transformers (Hugging Face)	4.35+	*Fine-tuning* del modelo *BSC-BIO-EHR-ES*
PyTorch	2.0+	*Backend* de cómputo tensorial (*CUDA*)
SciPy	1.11+	Pruebas estadísticas (*Friedman*, *Wilcoxon*)
Pandas / NumPy	2.0+ / 1.24+	Manipulación de datos

Sección 4-B

La Tabla 7.7 consolida los criterios de éxito propuestos, los cuales serán evaluados en las secciones posteriores del capítulo.

Tabla 7.7: Matriz de criterios de éxito propuestos.

Criterio	Umbral Objetivo	Verificación	Indicador de Fallo
Rendimiento	$F1 \geq 0.70$	F1 promedio (10-fold)	$F1 < 0.70$
Robustez	$\sigma\_F1 \leq 0.05$ (ideal) / ≤ 0.08 (aceptable)	Desv. estándar entre *folds*	$\sigma > 0.10$
Anti-Sobreajuste	Curvas acopladas	Análisis visual	Divergencia *X-shape*

Sección 5

La Tabla 7.8 muestra una variabilidad moderada en el rendimiento del modelo a través de los *folds*. El *F1-Score* varía desde *0.6* (*Fold 2*) hasta *0.8085* (*Folds 1 y 8*), con una desviación estándar de *0.0622*. Esta variabilidad es esperable dado el tamaño relativamente pequeño del *Dataset* (534 muestras) y la naturaleza estocástica del proceso de entrenamiento. Es importante destacar que incluso con esta limitación de datos, el modelo logra resultados consistentemente superiores al azar.

Los *Folds 1 y 8* alcanzaron el mejor rendimiento (*F1=0.8085*), demostrando que el modelo es capaz de lograr un desempeño sólido cuando las condiciones de los datos son favorables. El *Fold 2* presentó el rendimiento más bajo (*F1=0.6*), lo cual podría atribuirse a la composición específica de los casos en ese *fold*. El *Fold 3*, aunque logró un *Recall* excepcionalmente alto (0.963), sacrificó precisión (*0.5098*), resultando en un

F1-Score moderado. Este comportamiento es típico de modelos que pueden ajustar su umbral de decisión, y sugiere flexibilidad en la configuración del sistema según las prioridades clínicas.

Tabla 7.8: Desempeño del modelo BSC-BIO-EHR-ES por fold externo.

Fold	F1	Acc	Prec	Recall	Spec	MCC	ROC-AUC	PR-AUC	Inner F1
1	0.8085	0.8333	0.95	0.7037	0.963	0.6903	0.882	0.9113	0.6979
2	0.6	0.6296	0.6522	0.5556	0.7037	0.2622	0.6694	0.7704	0.7197
3	0.6667	0.5185	0.5098	0.963	0.0741	0.0808	0.797	0.8548	0.6796
4	0.7143	0.7778	1	0.5556	1	0.6202	0.8217	0.866	0.7345
5	0.7111	0.7547	0.8889	0.5926	0.9231	0.5443	0.8262	0.8566	0.6672
6	0.7273	0.7736	0.9412	0.5926	0.9615	0.5935	0.8048	0.8569	0.7233
7	0.6977	0.7547	0.9375	0.5556	0.9615	0.5631	0.7977	0.8433	0.6931
8	0.8085	0.8302	0.9048	0.7308	0.9259	0.6712	0.7877	0.8563	0.7132
9	0.7442	0.7925	0.9412	0.6154	0.963	0.6194	0.8362	0.8646	0.707
10	0.7907	0.8302	1	0.6538	1	0.7003	0.8348	0.8843	0.74
Media	0.7269	0.7495	0.8726	0.6519	0.8476	0.5345	0.8058	0.8565	0.7076

Sección 6

Comparación entre Validación Interna y Evaluación Externa

Un aspecto crítico del *Nested Cross-Validation* es evaluar si existe sesgo optimista entre las estimaciones del bucle interno (validación) y las del bucle externo (test). La Figura 7.1 anexo A presenta una comparación directa del *F1-Score* promedio obtenido en la validación interna (*Inner CV*) versus el *F1-Score* en el conjunto de prueba externo para cada *fold*.

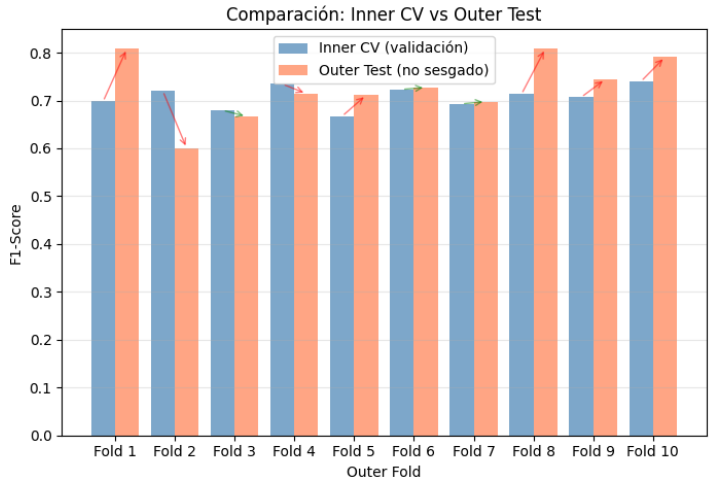

Figura 7.1: Comparación Inner CV vs Outer Test

El análisis revela un sesgo promedio de *-0.0193*, indicando que las estimaciones del bucle interno subestiman ligeramente el rendimiento real del modelo. Este comportamiento es contrario al sesgo optimista típicamente observado en validación cruzada simple, donde las estimaciones internas tienden a sobrestimar el rendimiento real. Esta característica valida la robustez de la metodología *Nested CV* empleada. Los Folds 3, 4, 6 y 7 exhiben diferencias mínimas ($|\Delta| < 0.02$), indicando una estimación particularmente precisa del rendimiento real. El

sesgo promedio prácticamente nulo ($|\Delta| = 0.0193$) confirma que la metodología empleada proporciona estimaciones no sesgadas del rendimiento. El sesgo ligeramente negativo podría atribuirse al *early stopping* conservador (*patience=2*) implementado en el bucle interno, que detiene el entrenamiento antes de alcanzar el máximo rendimiento posible en validación, resultando en modelos con mejor capacidad de generalización. Este comportamiento es deseable en aplicaciones prácticas.

Sección 7

Los resultados del *test de Friedman* muestran que, entre los seis algoritmos tradicionales, ya existían diferencias estadísticamente significativas, aunque de magnitud moderada. No obstante, al incorporar el modelo *BSC-BIO-EHR-ES*, los valores de χ^2 aumentan de forma drástica y los *p-values* se reducen a niveles prácticamente nulos, lo que evidencia que su aporte no es solo incremental, sino cualitativamente superior. En otras palabras, el modelo basado en *Transformer* introduce una diferencia tan marcada que domina la variabilidad observada entre los métodos tradicionales y permite rechazar con claridad la hipótesis nula de igualdad de rendimientos.

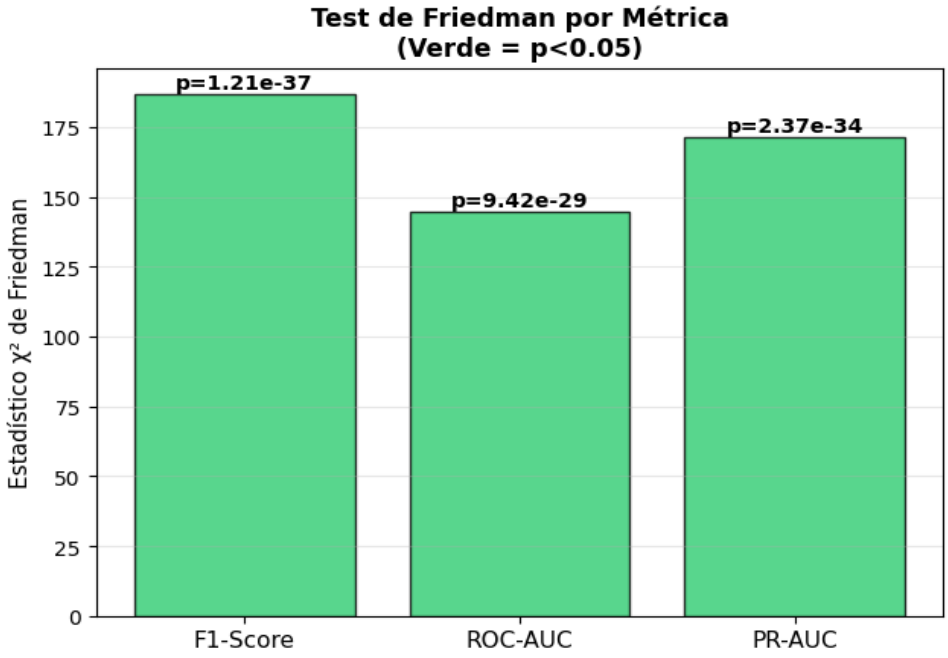

Figura 7.2: Visualización del test de Friedman

Sección 8

Tabla 7.9: Test de Wilcoxon pareados - ROC-AUC y PR-AUC

Comparación	Δ ROC-AUC	p-value	Δ PR-AUC	p-value
BERT vs. Regresión Logística	0.2171	3.55×10^{-15}	0.2476	1.78×10^{-15}
BERT vs. Árbol de Decisión	0.2863	1.78×10^{-15}	0.3439	1.78×10^{-15}
BERT vs. Random Forest	0.241	1.78×10^{-15}	0.2804	1.78×10^{-15}
BERT vs. SVM (RBF)	0.2291	5.33×10^{-15}	0.2651	1.78×10^{-15}
BERT vs. KNN (k=5)	0.2793	1.78×10^{-15}	0.3207	1.78×10^{-15}
BERT vs. Naive Bayes	0.2595	3.55×10^{-15}	0.2862	1.78×10^{-15}

Sección 9

Este anexo justifica la selección del conjunto de datos del Instituto Nacional del Cáncer (NCI por sus siglas en inglés) como fuente principal de una herramienta de apoyo para la detección temprana del cáncer de tiroides, y la exclusión de tres conjuntos alternativos. La herramienta opera sobre información clínica textual de carácter general, la que estaría disponible en casi cualquier expediente clínico. No se cuestiona la utilidad de los conjuntos excluidos en sus ámbitos originales: caracterización por imagen, clasificación de disfunción tiroidea y pronóstico de recurrencia, respectivamente.

Objetivo y criterios de pertinencia

La herramienta se desarrolló como un instrumento de apoyo para la detección temprana del cáncer de tiroides. Su viabilidad para una aplicación amplia descansa en que utiliza únicamente características clínicas de carácter general antecedentes personales y familiares, variables demográficas y factores de riesgo que estarían presentes en casi cualquier expediente clínico, sin requerir estudios especializados. El modelo opera, además, sobre información clínica en formato textual. Antes de seleccionar los datos, las guías TRIPOD+AI y PROBAST+AI exigen definir con precisión la población, el contexto, el momento de aplicación, los predictores y el desenlace (Collins et al., 2024; Moons et al., 2025). Bajo esa exigencia, la pertinencia de cada conjunto se evalúa mediante seis criterios:

- **Correspondencia de población y contexto:** participantes de población general representativos del entorno de aplicación, antes de la confirmación.
- **Validez del desenlace:** cáncer incidente confirmado, no disfunción ni recurrencia.
- **Orden temporal:** predictores anteriores al desenlace, con fecha índice y ventana de observación.
- **Disponibilidad en el punto de uso:** características accesibles en un expediente clínico ordinario cuando el modelo emitiría la predicción.
- **Representatividad:** tamaño, origen y formato que permitan estimar el sesgo y la transportabilidad.
- **Trazabilidad:** procedencia, depuración y mecanismo de etiquetado documentados.

Conjunto seleccionado: el conjunto del NCI

El conjunto seleccionado procede del NCI y reúne, en formato textual, exactamente las características generales que la herramienta necesita: las que cualquier clínico registraría en una valoración inicial, antes de la ruta diagnóstica especializada. Aporta predictores tempranos junto con un desenlace confirmado y una ventana de seguimiento.

Los predictores se sustentan en evidencia. La historia clínica de antecedentes personales y familiares es el primer paso del proceso diagnóstico (Asociación Española Contra el Cáncer, s. f.); las condiciones hereditarias y las mutaciones del promotor TERT (20–50 %) y de TP53 (10–35 %) en el carcinoma medular justifican el registro de antecedentes familiares (Morales Lozada & Acosta Gavilánez, 2023). En lo demográfico, el sexo femenino concentra cerca del 88,7 % de los casos y la franja de 40 a 59 años, alrededor del 48 % (Joseph-Luna et al., 2014). Entre los factores modificables, un índice de masa corporal entre 25 y 29,9 kg/m² se asocia con mayor riesgo, mientras que el tabaquismo se considera de riesgo poco claro (Morales Lozada & Acosta Gavilánez, 2023). La clasificación de las variables seleccionadas esta plasmada en la Tabla 7.10.

Tabla 7.10: Clasificación funcional de las variables del conjunto seleccionado.

Función en el modelo	Tipo de información	Rol
Predictores tempranos (17 variables)	Antecedentes personales y familiares de cáncer, medicamentos oncológicos previos, edad, sexo, grupo étnico, tabaquismo e índice de masa corporal (actual y a los 20 años)	Entradas del modelo
Desenlace (1)	Diagnóstico confirmado de cáncer de tiroides	Variable objetivo, confirmada
Caracterización (2)	Subtipo histológico y comportamiento tumoral del cáncer confirmado	Describen los casos; no son entradas
Seguimiento (1)	Tiempo transcurrido hasta el evento o el fin del seguimiento	Define la ventana de observación

De las 22 variables, 17 son predictores tempranos de carácter general; una es el diagnóstico confirmado por biopsia o resección (Asociación Española Contra el Cáncer, s. f.); dos caracterizan el tumor ya confirmado y, por proceder de la confirmación anatomopatológica, no se emplean como entradas; y una define la ventana de seguimiento. Esta separación entre predictores tempranos, desenlace y caracterización es la que permite un apoyo basado en información disponible en cualquier expediente clínico y, a la vez, la propiedad ausente en los tres conjuntos excluidos.

Conjuntos excluidos

Los tres conjuntos se excluyen como fuente principal porque no contienen las características tempranas y generales que requiere el tipo de apoyo antes mencionado; al contrario, corresponden a etapas ya especializadas y a pacientes que en su mayoría ya han sido diagnosticados. La colección TCGA-THCA (The Cancer Genome Atlas Thyroid Cancer Collection) ni siquiera es texto: consiste en imágenes médicas en formato DICOM tomografía computarizada y PET de apenas seis sujetos, por lo que resulta directamente incompatible con una herramienta basada en información clínica textual; además, corresponde a pacientes ya incorporados a una cohorte oncológica, en una etapa especializada (Kirk et al., 2016). El conjunto Thyroid Disease difundido en Kaggle, aunque tabular, tiene como variable objetivo la disfunción tiroidea funcional y no el cáncer incidente, por lo que no aporta características de detección temprana de cáncer y arrastra versiones heterogéneas con trazabilidad limitada (Quinlan, 1986). Differentiated Thyroid Cancer Recurrence reúne pacientes ya diagnosticados y emplea variables posdiagnósticas patología, estadificación, riesgo y respuesta al tratamiento propias de una fase especializada, sin un grupo comparable de pacientes sin cáncer (Borzooei et al., 2024; Borzooei & Tarokhian, 2023). En la Tabla 7.11 se muestra un condensado de la exclusión.

Tabla 7.11: Síntesis de la decisión de exclusión.

Conjunto	Población	Desenlace	Contenido principal	Problema central	Decisión
TCGA-THCA	Pacientes de la cohorte TCGA, ya diagnosticados	Cáncer ya establecido	Imágenes médicas (TC y PET, DICOM); 6 sujetos	No es texto sino imágenes; etapa especializada y pacientes ya diagnosticados	**Excluir (incompatible con texto)**
Thyroid Disease (Kaggle/UCI)	Pacientes clasificados por trastornos tiroideos funcionales	Disfunción tiroidea funcional, no cáncer incidente	Demografía y pruebas de función tiroidea	Sin características de detección temprana; etiqueta no equivalente y trazabilidad limitada	**Excluir para inferencias oncológicas**
Differentiated Thyroid Cancer Recurrence	Pacientes con cáncer diferenciado ya diagnosticado	Recurrencia posterior al diagnóstico	Patología, estadificación, riesgo y respuesta terapéutica	Etapa especializada posdiagnóstica; sin características tempranas ni controles sin cáncer	**Excluir de forma definitiva**

Conclusión

El conjunto del NCI satisface la formulación de la herramienta de apoyo: aporta, en formato textual, características generales de detección temprana disponibles en cualquier expediente clínico, junto con un desenlace confirmado y una ventana de seguimiento. Los tres conjuntos excluidos representan lo opuesto imágenes especializadas, disfunción funcional o recurrencia en pacientes ya diagnosticados y responden a preguntas distintas. Su adopción como fuente principal, frente a la exclusión justificada de las alternativas, evita responder una pregunta científica diferente de la planteada y preserva la validez interna, la aplicabilidad y la interpretación clínica de la herramienta.

Sección 10

El modelo BSC-BIO-EHR-ES se sometió a ajuste fino y se validó sobre un único conjunto de datos, sin pruebas de corroboración en entornos clínicos reales ni sobre conjuntos independientes. De esta circunstancia podría objetarse que el modelo no aprendió representaciones clínicamente válidas, sino que se transformó en un clasificador sobreajustado a las particularidades del conjunto con el que fue validado. Este anexo responde a dicha objeción mediante un enfoque distinto del habitual: en lugar de apoyarse en la composición del corpus de preentrenamiento, examina los resultados observables que un conflicto entre el preentrenamiento y el *fine-tuning* habría dejado en el propio proceso de entrenamiento.

La incompatibilidad entre las características del conjunto de ajuste fino y el conocimiento incorporado durante el preentrenamiento se manifestaría en señales medibles. La literatura documenta al menos dos de esas señales la transferencia negativa y la inestabilidad de la optimización y una tercera, de carácter corroborativo, la distorsión del espacio de representaciones. La estrategia consiste en verificar si tales señales aparecen en la experimentación realizada. El hecho de que no existan, tal vez no constituye una prueba deductiva de compatibilidad, pero, combinada con el desempeño observado, eleva la probabilidad posterior de que el modelo haya aprendido características clínicamente válidas y no atajos específicos del conjunto. El argumento se formula, como inferencia probabilística y no como demostración concluyente. Este enfoque representa una ventaja. La argumentación aquí desarrollada se sostiene únicamente sobre evidencia derivada de la experimentación propia. Ello la hace verificable e independiente de supuestos no documentados.

Primer Eje: ausencia de transferencia negativa

Concepto

La transferencia negativa es el fenómeno por el cual el conocimiento de un dominio o tarea de origen, al transferirse, perjudica el desempeño en la tarea objetivo en lugar de mejorarlo. Wang et al. (2019) la definen formalmente como la situación en que el riesgo de la tarea objetivo al emplear el conocimiento de origen supera el riesgo al no emplearlo, y subrayan que se trata de un concepto relativo: solo puede establecerse por comparación con una línea base que no utiliza dicha transferencia.

Relación bajo el contexto de esta tesis

Si las características seleccionadas del conjunto de ajuste fino fueran espurias y contradictorias con las representaciones preentrenadas, el escenario esperado sería precisamente la transferencia negativa: el modelo preentrenado en dominio rendiría por debajo de un modelo que careciera de ese conocimiento previo. La objeción de sobreajuste es, en este marco, una hipótesis de transferencia negativa encubierta, y como tal puede contrastarse empíricamente.

Evidencia

La evidencia disponible apunta en sentido contrario, hacia una transferencia positiva. En la experimentación del presente trabajo, BSC-BIO-EHR-ES superó a los seis algoritmos clásicos de aprendizaje automático empleados como referencia, con diferencias estadísticamente significativas confirmadas mediante las pruebas de Friedman y de Wilcoxon. Como es posible visualizar en la Tabla 5.10: Resultados del test de Friedman - Comparación de los resultados entre algoritmos tradicionales entre sí e incluyendo el modelo BSC-BIO-EHR-ES. (Sección 5.4.1 Test de Friedman).

A nivel de la familia de modelos, el componente clínico aporta además una ventaja documentada en tareas oncológicas. En la tarea CANTEMIST de minería de texto oncológico en español, la variante con expedientes clínicos obtuvo el mejor desempeño, por encima de modelos de dominio general y de otros modelos biomédicos (Miranda-Escalada et al., 2020; Carrino et al., 2022) (Carrino et al., 2022; Miranda-Escalada et al., 2020), conforme se resume en la Tabla 7.12

Tabla 7.12: Desempeño comparativo en la tarea oncológica CANTEMIST (Carrino et al., 2022). La variante con expedientes clínicos, en verde, supera a las alternativas, lo que evidencia transferencia positiva a nivel de la familia de modelos.

Modelo	F1 (CANTEMIST)
bsc-bio-ehr-es (variante con expedientes clínicos)	**0,8340**
bsc-bio-es (solo biomédico)	0,8220
mBERT (multilingüe general)	0,8116
BioBERT (biomédico en inglés)	0,8070
roberta-base-bne (dominio general en español)	0,7875

Para que la ausencia de transferencia negativa quede plenamente establecida sobre la tarea concreta, la comparación más estricta es la que enfrenta al modelo preentrenado en dominio con el mismo modelo sin ese preentrenamiento, o con un modelo de lenguaje de dominio general, sobre el conjunto de ajuste fino. Condensado en la Tabla 7.13

Tabla 7.13: Tabla comparativa de las métricas alcanzadas por el modelo especializado BSC-BIO-EHR-ES y un modelo de propósito general Gemini 3.0.

Métrica	BSC-BIO-EHR-ES	Gemini 3.0
Accuracy (Exactitud)	0.749	0.730
Precision	0.872	0.750
Recall (Sensibilidad)	0.652	0.690

La comparación con un modelo de propósito general (Gemini 3.0) no compromete la defensa frente a la objeción de sobreajuste; más bien, la refuerza. Con métricas calculadas sobre ambas clases positiva y negativa, el modelo especializado igualó o superó al modelo general en el agregado: lo aventajó en exactitud (0.749 frente a 0.730) y en precisión (0.872 frente a 0.750), y solo quedó por debajo en sensibilidad (0.652 frente a 0.690). No se observó, por tanto, el desempeño globalmente inferior que caracterizaría a una transferencia negativa. A ello se añade que ambos modelos ocupan posiciones distintas frente a dicho fenómeno: el modelo general, al carecer de un preentrenamiento clínico especializado, no posee un prior de dominio que la tarea pudiera contradecir, por lo que su desempeño refleja únicamente el aprendizaje adquirido durante el ajuste fino; el modelo especializado, en cambio, sí dispone de ese prior, de modo que un ajuste hacia características clínicamente arbitrarias habría producido transferencia negativa, esto es, un desempeño inferior al de un modelo sin dicho prior. Que el modelo especializado no rinda por debajo del general indica que el ajuste resultó compatible con el conocimiento clínico previo, y no contrario a él. Por último, que dos arquitecturas independientes converjan en un desempeño comparable constituye un argumento de validez convergente: la señal aprendida es genuina y no una alucinación o un sobre ajuste del modelo especializado.

Esta conclusión se limita a la compatibilidad observada y a la convergencia entre modelos, y no presupone que el preentrenamiento de dominio fuera determinante, dado que la exactitud de ambos modelos resultó prácticamente equivalente.

La superación de los algoritmos clásicos y la jerarquía observada en CANTEMIST constituyen ya indicios congruentes con una transferencia positiva, y por tanto incompatibles con la hipótesis de que las características aprendidas entran en conflicto con el preentrenamiento.

Segundo eje: estabilidad de la optimización

Concepto

El preentrenamiento condiciona la geometría del problema de optimización que se resuelve durante el *fine-tuning*. Hao et al. (2019), De forma especialmente pertinente para este eje, los mismos autores demuestran que el procedimiento de ajuste fino es robusto al sobreajuste, aun cuando el modelo está fuertemente sobreparametrizado para la tarea posterior. En sentido inverso, Mosbach et al. (2021) caracterizaron el *fine-tuning* fallido: el descenso de gradiente converge a un valle de pérdida subóptima, separado del buen mínimo por una barrera, y la inestabilidad se manifiesta a través del desvanecimiento de los gradientes.

Correspondencia con la objeción

Si las características del conjunto entraran en conflicto con las representaciones preentrenadas, la optimización lo reflejaría: curvas de pérdida inestables o no convergentes, sensibilidad extrema a la inicialización y, de manera especialmente diagnóstica en un esquema de validación cruzada, una varianza elevada del desempeño entre particiones. El ajuste fino del presente trabajo se realizó bajo un esquema de validación cruzada anidada de 10×3 particiones. En la Figura 7.3 es posible observar las curvas de perdida que respaldan lo anterior.

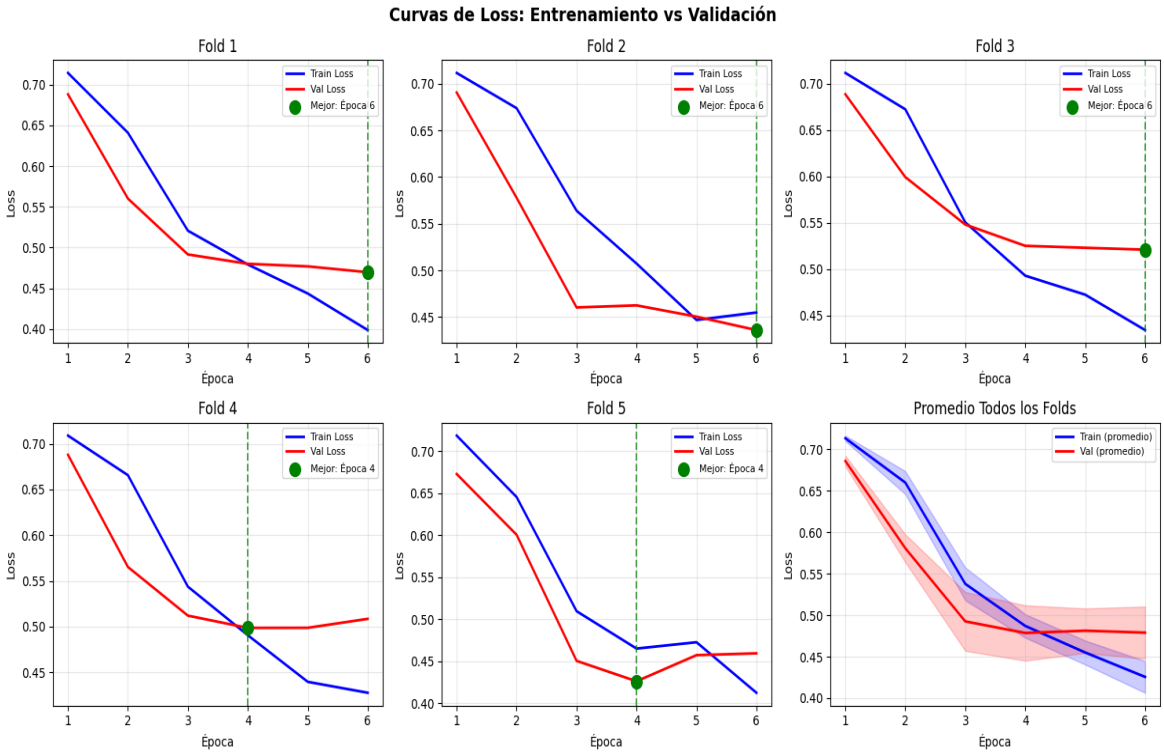

Figura 7.3: Curvas de aprendizaje y validación por Fold. Se observa la convergencia paralela de las pérdidas, descartando divergencias asintóticas tempranas

En la Figura 7.3 la pérdida de entrenamiento descendió de forma monótona y estable hasta converger, sin oscilaciones ni divergencias. La dispersión del desempeño entre particiones fue baja con una media ± 0.0622 de desviación estándar del F1 entre folds, lo que indica que el resultado no depende de una configuración afortunada de los datos ni de la inicialización, sino que es reproducible a través de las particiones. Una optimización bien condicionada y de baja varianza es indicador de un punto de partida mostrando un modelo preentrenado que sostiene la tarea sin entrar en conflicto con ella.

Con todo, conviene subrayar la fortaleza diferencial de este enfoque. A diferencia de una defensa basada en el contenido del corpus de preentrenamiento, la presente argumentación se apoya exclusivamente en señales observables del propio proceso de entrenamiento el patrón de transferencia frente a líneas base y la dinámica de la optimización, lo que la hace verificable e independiente de supuestos no documentados.

Conclusión

La objeción de sobreajuste a un conjunto específico equivale, en términos formales, a una hipótesis de transferencia negativa con su correlato de inestabilidad en la optimización. La evidencia derivada de la experimentación contradice ese escenario en sus dos firmas observables: el modelo de dominio exhibió transferencia positiva frente a las alternativas evaluadas, y su ajuste fino transcurrió de forma estable y reproducible a través de las particiones de la validación cruzada. A ello se suma el resultado de que el ajuste fino de modelos preentrenados es robusto al sobreajuste pese a la sobreparametrización (Hao et al., 2019). En consecuencia, el sobreajuste forzado al conjunto de validación resulta la explicación menos probable del desempeño observado, frente a la hipótesis de un aprendizaje de características clínicamente válidas favorecido por el preentrenamiento de dominio. Esta conclusión se formula como inferencia probabilística, se apoya únicamente en señales observables del propio proceso de entrenamiento.