ARTÍCULOS DE INVESTIGACIÓN

ANÁLISIS COMPARATIVO DE LA APLICACIÓN DE LOS ALGORITMOS DE APRENDIZAJE AUTOMÁTICO *RANDOM FOREST* Y REGRESIÓN LOGÍSTICA EN LA CLASIFICACIÓN DE CONJUNTOS DE TEXTOS

Stalin Francisco Calva Vicente
https://orcid.org/0009-0009-1163-0271
scalva3@utmachala.edu.ec
Universidad Técnica de Machala, Ecuador

Angel Ricardo Vera Santos
https://orcid.org/0000-0002-2300-6005
avera9@utmachala.edu.ec
Universidad Técnica de Machala, Ecuador

Freddy Anibal Jumbo Castillo
https://orcid.org/0000-0002-5200-7162
fjumbo@utmachala.edu.ec
Universidad Técnica de Machala, Ecuador

Johnny Paul Novillo Vicuña
https://orcid.org/0000-0002-4915-3441
jnovillo@utmachala.edu.ec
Universidad Técnica de Machala, Ecuador

Recibido: 26 de marzo del 2026 / Aceptado: 10 de junio del 2026
doi: https://doi.org/10.26439/interfases2026.n023.8710

RESUMEN. Gestionar la inmensa cantidad de datos no estructurados que generamos hoy sería impensable sin la clasificación automática de textos. Esta herramienta se ha vuelto indispensable para tareas tan variadas como el análisis de sentimientos o la detección de noticias falsas. Sin embargo, surge un dilema técnico importante: a pesar de la enorme oferta de algoritmos diseñados para estos fines, todavía no hay un consenso claro sobre cuál es el más efectivo para cada conjunto de datos específico. Para aliviar este problema, la presente investigación evalúa la robustez de dos modelos. Estos modelos son *random forest*, que funciona con conjuntos de árboles, y la regresión logística, un modelo lineal. El corpus empleado es Symptom2Disease, un conjunto de 1200 registros textuales con descripciones de síntomas en lenguaje natural distribuidos equitativamente entre 24 categorías diagnósticas que abarca patologías como psoriasis, diabetes, dengue y malaria.

Para este propósito, se ha adoptado un diseño experimental factorial con preprocesamiento estándar, la adición de aumento de datos para reducir la limitación de muestras y análisis de representación a través de N-gramas. Los resultados muestran que la regresión logística funciona mejor que *random forest*, pues alcanza un máximo de F1-Score de 0,9797 cuando se utiliza junto con el aumento de datos y los unigramas. Para textos médicos cortos, la linealidad del modelo y el enriquecimiento sintético del corpus constituyen una solución más eficiente y precisa en comparación con los conjuntos no lineales.

PALABRAS CLAVE: aprendizaje automático / procesamiento del lenguaje natural / clasificación de textos clínicos / regresión logística / *random forest*

COMPARATIVE ANALYSIS OF THE APPLICATION OF THE RANDOM FOREST AND LOGISTIC REGRESSION MACHINE LEARNING ALGORITHMS FOR TEXT CORPUS CLASSIFICATION

ABSTRACT. Managing the vast amount of unstructured data we generate today would be unthinkable without automatic text classification. This tool has become indispensable for tasks as varied as sentiment analysis and combating fake news. However, a significant technical dilemma arises: despite the vast array of algorithms designed for these purposes, there is still no clear consensus on which is the most effective for each specific dataset. To address this issue, the present study compares the robustness of two models: Random Forest, which operates using a set of trees, and Logistic Regression, a linear model. The corpus used is Symptom2Disease, a dataset of 1,200 text records containing descriptions of symptoms in natural language, evenly distributed across 24 diagnostic categories, covering conditions such as psoriasis, diabetes, dengue, and malaria. For this purpose, a factorial experimental design was adopted, incorporating standard preprocessing, data augmentation to mitigate sample size limitations, and representation analysis via N-grams. In other words, it appears that Logistic Regression performs better than Random Forest, achieving a maximum F1 score of 0.9797 when used in conjunction with data augmentation and unigrams. For short medical texts, the linearity of the model and synthetic corpus enrichment prove to be a more efficient and accurate solution compared to non-linear models.

KEYWORDS: machine learning / natural language processing / clinical text classification / logistic regression / random forest

INTRODUCCIÓN

La digitalización ha disparado la creación de textos, lo que ha alterado profundamente los mecanismos de gestión de información. Hoy, la prioridad no es solo generar datos, sino entender cómo esa marea de contenido impacta en la toma de decisiones y en el uso de la información disponible (Kowsari et al., 2019). Hablamos de un ecosistema masivo de información no estructurada que abarca desde artículos de prensa y redes sociales hasta registros médicos o legales cuyo valor real es inmenso. Sin embargo, ese potencial solo se puede aprovechar si implementamos procesos de análisis y tratamiento de datos que sean realmente consistentes y rigurosos. Dado el punto anterior, la clasificación automática de textos, un tema central del procesamiento de lenguaje natural (NLP), representa una de sus tareas básicas para dar sentido a grandes cantidades de datos (Aggarwal & Zhai, 2012). Su aplicabilidad es transversal a múltiples dominios, lo que facilita desde la detección de noticias falsas (Báez et al., 2022) hasta el análisis de sentimientos (positivo, negativo, neutral) de un texto (Calero Sánchez et al., 2024). Ello evidencia el enorme impacto de la inteligencia artificial en el aprendizaje automático aplicado al procesamiento de texto.

A pesar de los avances logrados y del amplio abanico de algoritmos disponibles, persiste un problema central: la complejidad de elegir el modelo que mejor se adapte a un objetivo o *dataset* específico (Li et al., 2022). Desde modelos lineales tradicionales hasta conjuntos de algoritmos complejos, la literatura existente detalla diferencias significativas en el rendimiento en relación con el tamaño de los datos, la dimensión del corpus y el tipo de lenguaje (Cardoso et al., 2023). Tal variabilidad resalta la importancia de enfoques comparativos sólidos que guíen a los investigadores y profesionales en la selección de la mejor herramienta de utilidad para sus respectivos dominios.

En el ámbito clínico, la clasificación automática de textos médicos enfrenta desafíos adicionales en comparación con otros dominios de texto. La alta especificidad del vocabulario médico, la presencia de términos con significados variables según el contexto sintomático y la frecuente escasez de corpus etiquetados de gran escala condicionan tanto la elección del algoritmo como la estrategia de representación vectorial adoptada (Báez et al., 2022). A esto se suma que, en entornos de apoyo al diagnóstico, la interpretabilidad del modelo cobra un valor estratégico: los profesionales de la salud no solo requieren predicciones precisas, sino también comprensión de los factores que motivaron una clasificación determinada. Esta exigencia convierte la elección del algoritmo en una decisión que equilibra simultáneamente exactitud, eficiencia computacional e interpretabilidad clínica, tres dimensiones en las que los modelos clásicos de aprendizaje automático se distinguen de manera diferenciada.

La elección de *random forest* y regresión logística como objetos de comparación responde a una estrategia metodológica deliberada: ambos representan paradigmas opuestos en la taxonomía del aprendizaje supervisado. Uno es lineal y probabilístico,

el otro es no lineal y se basa en conjuntos de árboles. Ambos figuran entre los clasificadores de referencia más ampliamente utilizados en la literatura de clasificación de textos médicos (Pranckevičius & Marcinkevičius, 2017). Al contrastar específicamente estos dos enfoques, la investigación busca aportar evidencia empírica sobre cuál paradigma resulta más adecuado para el dominio de síntomas clínicos en lenguaje natural, una pregunta que, a pesar de su relevancia práctica, carece de respuesta sistemática en el contexto de corpus médicos de tamaño pequeño y equilibrado.

Para mitigar este problema, esta investigación realiza un análisis comparativo de los rendimientos de dos algoritmos de clasificación bien aceptados, basados en diferentes fundamentos: *random forest* y regresión logística. *Random forest* es un algoritmo de conjunto bien conocido que construye múltiples árboles de decisión y es altamente capaz de acomodar las necesidades de datos de alta dimensión y evita el sobreajuste en grandes conjuntos de datos (Pal, 2005). Sin embargo, la regresión logística es un modelo lineal probabilístico que es favorecido debido a su simplicidad, eficiencia computacional e interpretabilidad, y descrito como un modelo de referencia confiable en muchos de los problemas de clasificación de textos (Genkin et al., 2007).

La investigación adopta un método experimental basado en pruebas controladas y exhaustivas. Más allá del preprocesamiento tradicional (tokenización, lematización), se evalúan dos estrategias de optimización complementarias: el aumento del léxico del corpus y la implementación de N-gramas para la vectorización por frecuencia de término-frecuencia inversa de documento (TF-IDF), un método estadístico que pondera la relevancia de cada palabra según su capacidad discriminativa en el texto, con la intención de capturar el contexto sintáctico de los síntomas. Para ello, se crearon múltiples escenarios de prueba para separar el efecto de estas variables en el rendimiento de *random forest* y de regresión logística. Luego, cada modelo fue entrenado y se compararon ambos estadísticamente sobre la base de los siguientes indicadores establecidos en la industria y la academia: precisión y *recall*, así como matriz de confusión y F1 *score*. Con ello, se pudo proporcionar una evaluación objetiva y multidimensional sobre si ambos modelos tuvieron éxito en la tarea de clasificación.

En cuanto a trabajos relacionados, investigaciones recientes han explorado la clasificación de textos médicos y la predicción de enfermedades a partir de descripciones sintomáticas en lenguaje natural comparando modelos clásicos y avanzados. Hassan et al. (2024) aplicaron modelos de lenguaje y aprendizaje profundo sobre el mismo corpus Symptom2Disease utilizado en esta investigación, y lograron una exactitud del 99,58 % con el modelo de normalización de conceptos médicos-representaciones de codificador bidireccional de transformadores (MCN-BERT); no obstante, dicho desempeño se obtiene a expensas de una complejidad computacional significativamente mayor frente a los enfoques clásicos aquí propuestos. Ello refuerza la utilidad de los modelos más simples e interpretables en escenarios de recursos limitados.

Por su parte, Al-qarni & Algarni (2025) compararon técnicas de aprendizaje profundo (LSTM, CNN-LSTM, GRU) frente a métodos clásicos basados en TF-IDF para la predicción de enfermedades a partir de texto libre. Concluyeron que los clasificadores lineales tradicionales mantienen ventajas en interpretabilidad y eficiencia computacional cuando el volumen de datos es reducido. En la misma línea, Vyas et al. (2025) evaluaron el sistema SympTextML para la predicción multiclase de enfermedades mediante descripciones de síntomas en lenguaje natural empleando representaciones TF-IDF sobre 24 categorías diagnósticas, y hallaron que los modelos clásicos bien configurados pueden competir con arquitecturas más complejas en corpus de tamaño reducido. Estas investigaciones subrayan la pertinencia de comparar los algoritmos de regresión logística y *random forest* en el dominio clínico, especialmente en tareas en las que la interpretabilidad y la eficiencia constituyen criterios de selección tan determinantes como la exactitud de la clasificación.

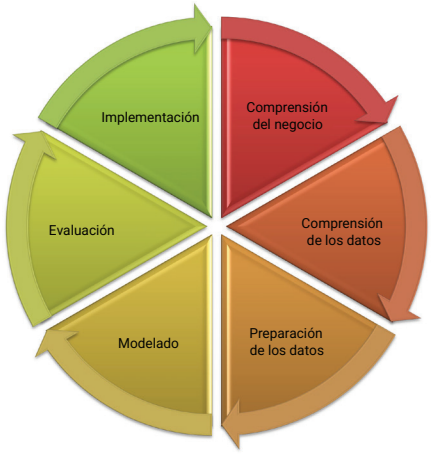
METODOLOGÍA

El estudio actual está diseñado bajo un marco cuantitativo, comparativo y experimental que evalúa el rendimiento de los algoritmos de aprendizaje automático *random forest* y regresión logística en la clasificación de textos en el dominio de síntomas y enfermedades.

En el desarrollo metodológico, adoptamos el modelo CRISP-DM (proceso estándar interindustrial para la minería de datos), el cual describe el marco de seis etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

Figura 1

Fases del modelo de referencia CRISP-DM



Nota. Los datos proceden de *CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc*, por P. Chapman, J. Clinton, R. Kerber, T. Reinartz, C. Shearer y R. Wirth, 2000, DaimlerChrysler (https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf).

Comprensión del negocio

El análisis automatizado de la información médica es un medio importante para mejorar los procesos de diagnóstico y la gestión del conocimiento clínico en la era digital. Con eso en mente, el estudio tiene como objetivo verificar si los algoritmos clasificadores *random forest* o regresión logística funcionan efectivamente en un texto médico que detalla síntomas y enfermedades. Además, busca proporcionar evidencia empírica que optimizará el análisis del sistema y la previsión en futuras investigaciones de salud digital.

En un sentido más amplio, el contexto de aplicación comprende sistemas de apoyo a la decisión clínica (*clinical decision support systems*, CDSS) que procesen narrativas escritas por pacientes en plataformas de telemedicina, portales de salud electrónica y sistemas de triaje remoto. La capacidad de clasificar automáticamente estas descripciones textuales en una categoría diagnóstica permitirá orientar al paciente hacia el especialista correspondiente, reducir tiempos de espera y optimizar la asignación de recursos en atención primaria. La validación experimental de modelos eficientes e interpretables para este fin constituye, por tanto, un aporte concreto a la infraestructura técnica de la salud digital (Calero Sánchez et al., 2024).

Comprensión de los datos

Para el desarrollo experimental, se utilizó el conjunto de datos Symptom2Disease, obtenido del repositorio público de ciencia de datos Kaggle y publicado originalmente por Barman et al. (2023). Este corpus fue organizado utilizando un formato estructurado, elegido por ser adecuado para la tarea de procesamiento de lenguaje natural involucrado en el dominio médico. Se compone de 1200 registros estructurados en dos columnas esenciales que caracterizan las variables de entrada y salida del modelo:

- label (variable dependiente): tipo de etiqueta categórica que representa el diagnóstico patológico. El conjunto cubre un total de 24 enfermedades como psoriasis, diabetes, dengue y COVID-19, etcétera.
- text (variable independiente): la descripción en lenguaje natural de los síntomas reportados por el paciente asociados con la enfermedad.

A modo de ilustración, la Tabla 1 presenta dos registros representativos del corpus que muestran la estructura de ambas variables tal y como fueron procesadas durante el experimento:

Tabla 1

Ejemplos representativos del corpus Symptom2Disease

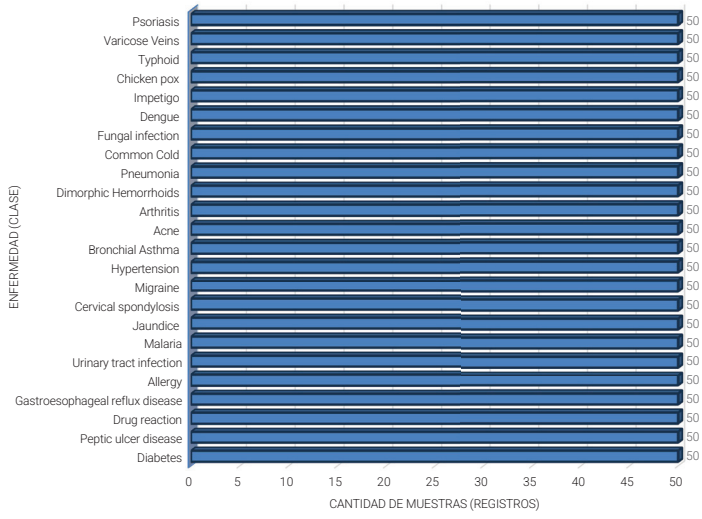
ID	label (diagnóstico)	text (descripción de síntomas)
0	Psoriasis	I have been experiencing a skin rash on my arms, legs, and torso for the past few weeks. It is red, itchy, and covered in dry, scaly patches.
600	Diabetes	I've been feeling really thirsty lately and urinating more than usual. I've also been losing weight unexpectedly. My vision has become blurry and I feel tired all the time.

Nota. Cada registro consta de una etiqueta de diagnóstico y una descripción de síntomas en inglés, en lenguaje coloquial. Los ejemplos proceden de *Symptom2Disease: Diseases and Natural Language Symptom Descriptions*, por N. Barman, F. Karim y K. Sharma, 2023.

Una clave para dar sentido a los datos es probar la distribución de las clases, ya que un desequilibrio puede sesgar el modelo hacia la mayoría de las clases. Como se demuestra en la Figura 2, el análisis exploratorio de datos verifica que los estos están equilibrados con 50 instancias de cada una de las 24 enfermedades con números correspondientes. Esta distribución uniforme también permite que los modelos de aprendizaje automático (*random forest* y regresión logística) aprendan los patrones típicos de cada patología de manera similar, por lo que no deberían requerir el uso inicial de métodos de submuestreo o sobremuestreo.

Figura 2

Distribución de muestras por enfermedad en el conjunto de datos

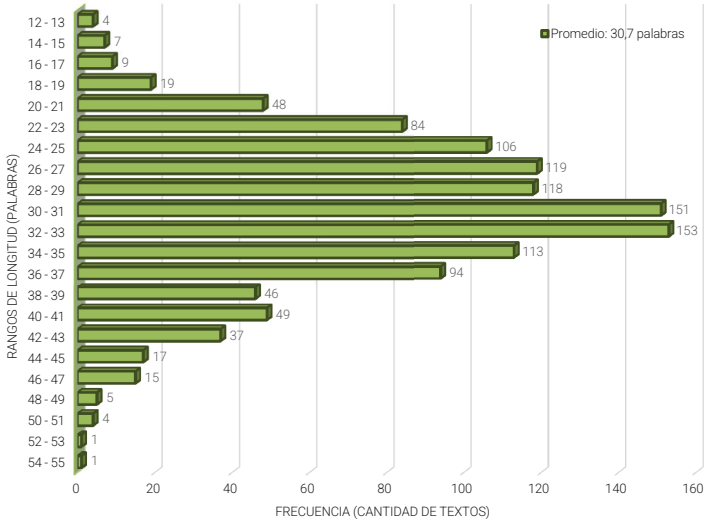


Nota. Se observa una distribución equitativa con 50 registros por clase, lo que asegura un entrenamiento balanceado. Los datos proceden de *Symptom2Disease: Diseases and Natural Language Symptom Descriptions*, por N. Barman, F. Karim y K. Sharma, 2023

Los textos también fueron analizados para determinar la longitud de las descripciones de los síntomas. Del histograma en la Figura 3, la longitud de los textos es aproximadamente igual a la normal (el promedio es de 30,7 palabras). Esto sugiere que los modelos funcionarían con textos cortos y directos, y la presencia de términos importantes particulares tendría un alto peso discriminativo, por lo que se justifica la exploración en representación basada en unigramas.

Figura 3

Distribución de la longitud de los textos (conteo de palabras)



Nota. Distribución del conteo de palabras por registro, con un promedio de 30,7 palabras por descripción. El volumen principal de los textos clínicos se concentra entre las 20 y 40 palabras. Los datos proceden de *Symptom2Disease: Diseases and Natural Language Symptom Descriptions*, por N. Barman, F. Karim y K. Sharma, 2023.

Finalmente, para caracterizar el dominio semántico del corpus, se generó una nube de palabras (WordCloud) que visualiza los términos más frecuentes. La Figura 4 destaca palabras como *skin* (piel), *pain* (dolor), *fever* (fiebre), *cough* (tos) y *discomfort* (malestar). La predominancia de estos sustantivos confirma que el vocabulario es altamente específico del dominio clínico sintomatológico, lo cual es ideal para el entrenamiento de clasificadores supervisados.

Figura 4

Nube de palabras de los términos médicos más frecuentes

Preparación de los datos

El corpus textual se preprocesó para lograr datos de alta calidad y uniformes antes del modelado. Las principales tareas trabajadas se hicieron de acuerdo con los principios de la minería de texto estándar presentados por Manning et al. (2008):

- Limpieza de datos: eliminar duplicados, caracteres nulos e irrelevantes.
- Unificación textual: todo el texto se ha reducido a minúsculas y se han eliminado los signos de puntuación.
- Tokenización: segmentación de los síntomas en palabras clave.
- Eliminación de stopwords: eliminar palabras sin semántica.
- Lematización: la normalización lingüística de los términos a su forma inicial.
- Aumento de datos (data augmentation): también utilizamos el método de intercambio aleatorio de palabras (random swap) con los datos de entrenamiento. Este enfoque está respaldado por Wei & Zou (2019) en su marco easy data augmentation (EDA), en el cual los textos originales se generan y utilizan como variantes sintéticas para hacer el modelo más sólido y manejar el posible sobreajuste que surge de su falta de datos por clase.
- Preservación del significado clínico: siguiendo la advertencia metodológica de Wei & Zou (2019), quienes señalan que el intercambio excesivo de palabras puede comprometer la validez de las etiquetas originales en textos breves, en el presente estudio se constató que el random swap preserva razonablemente

el significado clínico de cada registro. La Tabla 2 muestra ejemplos representativos del proceso de aumento, en el que se ilustra cómo el intercambio de posición de dos palabras mantiene el cuadro sintomático identificable.

Tabla 2

Ejemplos de textos originales y sus versiones aumentadas mediante random swap

Tipo	label	text
Original	Psoriasis	I have been experiencing a *skin rash* on my arms, legs, and torso. It is red, itchy, and covered in dry, scaly patches.
Aumentado (*random swap*)	Psoriasis	I have been experiencing a *rash skin* on my arms, legs, and torso. It is red, itchy, and covered in scaly, dry patches.

Nota. El intercambio de posición de dos palabras (señaladas en cursiva) no altera el diagnóstico subyacente ni elimina los términos clínicos clave. El corpus aumentado conserva la misma etiqueta original.

- Vectorización y representación: transformación de textos utilizando el esquema de frecuencia de término-inversa frecuencia de documento (TF-IDF). Este enfoque, que ha sido bien fundamentado en la literatura de clasificación de textos (Alsaeedi, 2020), pondera la relevancia de cada síntoma al reducir el peso de los términos no representativos que no expresan poderes discriminativos y, por lo tanto, para modelos lineales y de conjunto, proporciona el ajuste del modelo de entrada.

El diseño del *pipeline* de preprocesamiento respondió a las particularidades lingüísticas del corpus. La lematización desempeña un papel crítico en textos clínicos, pues permite unificar variantes morfológicas de términos diagnósticos: palabras como *itching*, *itches* e *itchy* se reducen al lema "itch", lo que disminuye la dispersión del espacio vectorial sin pérdida semántica. La eliminación de *stopwords* actúa como filtro de ruido léxico, priorizando los sustantivos y adjetivos que constituyen el núcleo de cada descripción sintomática (Manning et al., 2008).

En cuanto a la vectorización TF-IDF, su ventaja sobre una representación simple de frecuencias radica en ponderar los términos según su relevancia discriminativa en el corpus: un síntoma infrecuente, pero diagnósticamente determinante como *petechiae*, asociado al dengue hemorrágico, recibe un peso muy superior al de términos genéricos como *feeling* que aparecen en múltiples categorías. Esta propiedad es especialmente valiosa en corpus médicos, en los que el vocabulario específico de cada patología constituye el principal marcador de identidad diagnóstica (Alsaeedi, 2020).

Modelado

En esta sección, el proceso de modelado se ha realizado con la aplicación de dos algoritmos típicos de aprendizaje supervisado: regresión logística y *random forest*. La elección de estos modelos es una cuestión metodológica basada en su diferente enfoque, complejidad y literatura científica respecto a la clasificación de textos.

La regresión logística fue seleccionada como un modelo lineal probabilístico prevalente en clasificaciones binarias y multiclase en el que se puede interpretar el efecto de las variables independientes en la predicción final. Su bajo costo computacional, facilidad de implementación y adecuación en datos de texto vectorizados por TF-IDF lo convierten en un punto de referencia confiable para comparar modelos más sofisticados, según reportan Pranckevičius y Marcinkevičius (2017). Los autores han demostrado que el modelo produce con alta precisión en corpus de reseñas y textos cortos, incluso ha superado a clasificadores más grandes en estabilidad.

Desde una perspectiva matemática, en su extensión multiclase, la regresión logística opera bajo el esquema multinomial, en el que, para cada clase k, se estima la probabilidad condicional $P(y = k|x)$ mediante la función *softmax* aplicada a la combinación lineal de los vectores TF-IDF de entrada. El parámetro de regularización C controla la penalización sobre los coeficientes del modelo: valores bajos implican regularización fuerte, mientras que valores altos permiten mayor ajuste a los datos de entrenamiento. Con $C = 1$, la regularización L2 equilibra la capacidad de ajuste con la generalización, lo que previene el sobreajuste en espacios de alta dimensionalidad como los vectores TF-IDF, cuya cardinalidad puede superar los diez mil atributos activos (Genkin et al., 2007). Esta propiedad resulta determinante en un corpus de texto médico breve, en el que la mayoría de los términos del vocabulario son infrecuentes y el riesgo de colinealidad entre características es elevado.

Por el contrario, el *random forest* fue elegido como modelo de tipo conjunto que se construye alrededor de árboles de decisión, ya que es un método ideal para alta dimensionalidad, interacción no lineal y evasión de sobreajuste. Según Sun et al. (2020), la integración de múltiples árboles de decisión en un ensamble proporciona mayor estabilidad predictiva y mejora la generalización frente a datos complejos y no estructurados, lo cual es deseable para reducir la dispersión presente en el procesamiento de texto no estructurado.

Para proporcionar una evaluación sistemática del efecto de los métodos probados, se ha establecido un diseño factorial 2 x 2 x 2 que conduce a 8 experimentos controlados. Estas situaciones permiten determinar el alcance del efecto de los algoritmos, el enfoque de aumento de datos y la representación de N-gramas (Tabla 3).

Tabla 3

Diseño experimental

N.°	Modelo	Técnica de datos	Representación (TF-IDF)	Hipótesis/objetivo del experimento
1	Regresión logística	Originales	Unigramas (1-gram)	Línea base (*baseline*): rendimiento estándar sin mejoras
2	Regresión logística	Originales	N-gramas (1,2)	Evaluar si capturar frases como *dolor de pecho* mejora la detección con datos limitados
3	Regresión logística	*Data augmentation*	Unigramas (1-gram)	Evaluar si la variabilidad léxica (sinónimos) compensa la falta de contexto
4	Regresión logística	*Data augmentation*	N-gramas (1,2)	Modelo óptimo lineal: combinación de riqueza de datos y contexto semántico
5	*Random forest*	Originales	Unigramas (1-gram)	Línea base (ensamble): evaluar capacidad base del modelo no lineal
6	*Random forest*	Originales	N-gramas (1,2)	Verificar si aumentar la dimensionalidad (más *features* por los bigramas) afecta al RF
7	*Random forest*	*Data augmentation*	Unigramas (1-gram)	Evaluar la reducción de sobreajuste (*overfitting*) mediante más datos
8	*Random forest*	*Data augmentation*	N-gramas (1,2)	Modelo óptimo ensamble: prueba de estrés con máxima complejidad de datos y *features*

Nota. Resumen de los escenarios para el análisis comparativo, variando el algoritmo, la técnica de aumento de datos y la tokenización.

No se consideraron otros algoritmos, incluidos los SVM y los modelos de aprendizaje profundo, ya que el objetivo no es analizar sistemáticamente todos los métodos existentes, sino comparar dos enfoques representativos y contrastantes: la regresión logística (lineal y explicativa) y un modelo de conjunto no lineal (*random forest*). De ese modo, esta metodología permite obtener hallazgos interpretables y equilibrados sobre cómo la naturaleza del modelo afecta la clasificación de textos médicos.

Con el propósito de garantizar la reproducibilidad de los experimentos, la Tabla 4 detalla la configuración de los hiperparámetros empleados en cada componente del *pipeline*: los dos clasificadores y el vectorizador TF-IDF. Los valores indicados corresponden a los utilizados en la implementación experimental, ya sea que hayan sido especificados explícitamente en el código o correspondan al valor predeterminado de la biblioteca scikit-learn.

Tabla 4

Configuración de hiperparámetros de los modelos y el vectorizador TF-IDF

Modelo	Hiperparámetro	Valor utilizado
Regresión logística	solver	lbfgs
Regresión logística	C (regularización L2)	1,0 (predeterminado)
Regresión logística	max_iter	1000
Regresión logística	multi_class	auto (predeterminado)
Regresión logística	random_state	42
Random forest	n_estimators	100
Random forest	max_depth	none (sin límite)
Random forest	min_samples_split	2 (predeterminado)
Random forest	criterion	gini (predeterminado)
Random forest	random_state	42
TF-IDF Vectorizer	ngram_range	(1,1) o (1,2) según experimento
TF-IDF Vectorizer	max_features	none (sin límite)
TF-IDF Vectorizer	min_df	1 (predeterminado)

Nota. Los valores marcados como *(predeterminado)* corresponden a los valores por defecto de scikit-learn y fueron mantenidos sin modificación. La semilla aleatoria random_state = 42 se fijó en todos los componentes para garantizar la reproducibilidad de los experimentos. El hiperparámetro ngram_range varía según el escenario experimental definido en la Tabla 3.

Los dos modelos se entrenaron con las mismas configuraciones experimentales. Además, para evitar sesgos en la estimación del error de generalización, se llevó a cabo una validación cruzada estratificada (Stratified K-Fold con $k = 5$). Esta estrategia, alineada con las recomendaciones de Vabalas et al. (2019) para la validación de algoritmos con tamaños de muestra pequeños, mantiene la misma proporción de las clases en cada partición, lo que minimiza efectivamente el sesgo de selección y proporciona criterios de evaluación sólidos. Es fundamental precisar que el proceso de *data augmentation* fue aplicado exclusivamente sobre el conjunto de entrenamiento dentro de cada partición (*fold*), es decir, los textos sintéticos fueron generados a partir de los datos de entrenamiento de ese *fold* después de realizar la división, y el vectorizador TF-IDF fue ajustado únicamente sobre dichos datos aumentados. El conjunto de validación de cada *fold* permaneció intacto y no fue expuesto durante el proceso de aumento ni durante el ajuste del vectorizador, lo que garantizó que no se produzca filtración de datos (*data leakage*) que pudiera inflar artificialmente las métricas de desempeño reportadas.

Evaluación

Siguiendo a Powers (2008), el rendimiento de cada modelo se evalúa considerando métricas estándar en el aprendizaje supervisado para la evaluación de clasificadores en escenarios equilibrados:

- Precisión: proporción de predicciones correctas entre los casos positivos estimados
- Exhaustividad (recall): proporción de casos positivos correctamente identificados
- Puntuación F1: media armónica entre precisión y recall para balances de rendimiento
- Matriz de confusión: análisis de aciertos y errores por clase

Estas métricas permiten determinar qué algoritmo realiza un mejor trabajo en lograr una clasificación bien equilibrada de textos médicos.

Implementación

El experimento utilizó el entorno de programación Python con pandas, numpy, scikit-learn y nltk como las bibliotecas personalizadas para el procesamiento del lenguaje natural y modelado de los datos.

Los resultados establecen las posibles conclusiones sobre la aplicabilidad de cada modelo en el campo del análisis automatizado de información médica, lo que proporciona información útil para futuras investigaciones en clasificación de textos.

RESULTADOS

Para evaluar la eficiencia de los modelos propuestos, se llevaron a cabo los ocho escenarios experimentales especificados en la metodología utilizando validación cruzada estratificada (Stratified K-Fold) de $k = 5$. A continuación, el rendimiento de los algoritmos se presenta, brevemente, en términos de diferentes configuraciones de procesamiento de datos y representación vectorial.

Desempeño del modelo de regresión logística

En primera instancia, se evaluó la capacidad del modelo lineal utilizando el conjunto de datos original sin técnicas de generación sintética. Los resultados se detallan en la Tabla 5.

Tabla 5

Desempeño del modelo de regresión logística con conjuntos de datos originales

N.°	Modelo	Técnica de datos	Representación (TF-IDF)	Precisión	*Recall*	F1-Score
1	Regresión logística	Originales	Unigramas (1-gram)	0,9728	0,9700	0,9695
2	Regresión logística	Originales	N-gramas (1,2)	0,9705	0,9667	0,9663
		Promedio		0,9716	0,9684	0,9679

Nota. Evaluación de la línea base (*baseline*) utilizando representaciones de unigramas y N-gramas sin técnicas de aumento de datos.

Como se observa en la Tabla 5, la regresión logística establece una línea base muy alta incluso sin aumento de datos: alcanza un F1-Score promedio de 0,9679. Al comparar las representaciones, el modelo basado en unigramas (Exp. 1) obtuvo un rendimiento ligeramente superior (0,9695) frente al de bigramas (0,9663), lo que sugiere que, con los datos originales, la inclusión de pares de palabras no aportó una ganancia significativa, incluso introdujo un leve ruido en la clasificación.

Posteriormente, se evaluó el impacto de la técnica de *data augmentation* en el modelo lineal. Los resultados de esta segunda fase se presentan en la Tabla 6.

Tabla 6

Impacto del data augmentation *en el modelo de regresión logística*

N.°	Modelo	Técnica de datos	Representación (TF-IDF)	Precisión	*Recall*	F1-*Score*
3	Regresión logística	*Data augmentation*	Unigramas (1-gram)	0,9823	0,9800	0,9797
4	Regresión logística	*Data augmentation*	N-gramas (1,2)	0,9765	0,9733	0,9731
		Promedio		0,9794	0,9766	0,9764

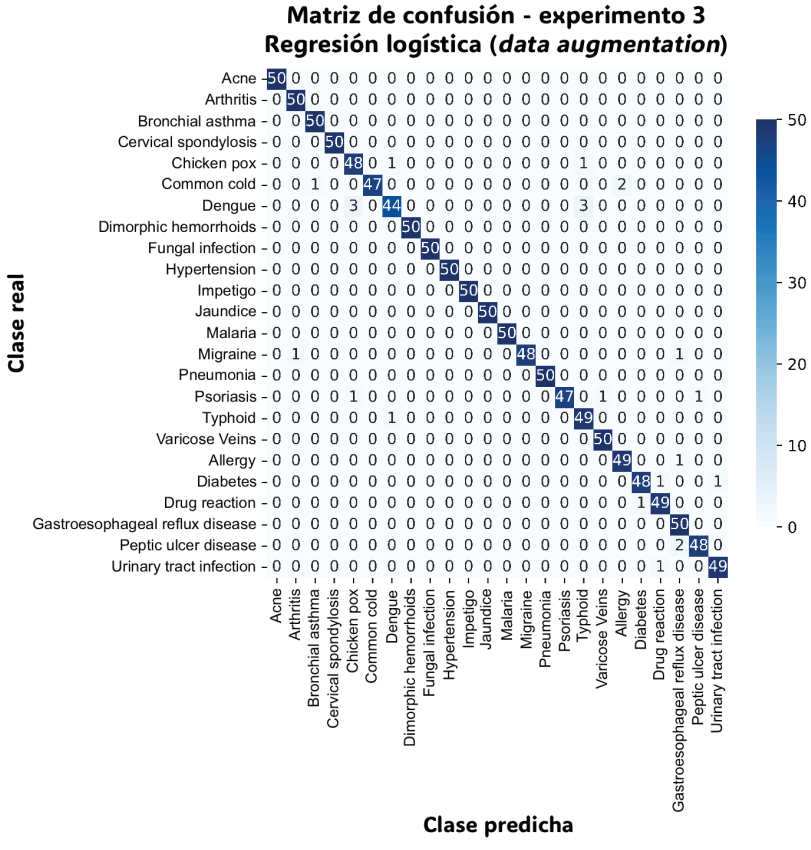
Nota. Comparación de métricas tras duplicar el corpus de entrenamiento mediante la técnica de intercambio aleatorio de palabras (*random swap*).

La Tabla 6 evidencia que la incorporación de datos sintéticos mejoró el rendimiento global del modelo: se elevó el F1-Score promedio a 0,9764. Es destacable que el experimento 3 (*data augmentation* con unigramas) alcanzó el desempeño máximo de todo el estudio con una precisión del 98,23 % y un F1-Score de 0,9797. Esto confirma que enriquecer la variabilidad léxica mediante el intercambio de palabras potencia la capacidad discriminativa del modelo lineal más que el aumento de la complejidad semántica (N-gramas).

Para ilustrar la precisión de este escenario óptimo, se presenta la matriz de confusión del experimento 3 (Figura 5).

Figura 5

Matriz de confusión del mejor modelo de regresión logística (Exp. 3)



Nota. Análisis de errores para la configuración óptima con *data augmentation* y unigramas, muestra la distribución de predicciones correctas (diagonal) y falsos positivos.

Desempeño del modelo *random forest*

A continuación, se presenta el análisis del comportamiento del algoritmo de ensamble (*random forest*). Primero, se evaluó su desempeño con los datos originales, tal como se muestra en la Tabla 7.

Tabla 7

Desempeño del modelo random forest *con conjuntos de datos originales*

N.°	Modelo	Técnica de datos	Representación (TF-IDF)	Precisión	*Recall*	F1-Score
5	*Random forest*	Originales	Unigramas (1-gram)	0,9594	0,9567	0,9555
6	*Random forest*	Originales	N-gramas (1,2)	0,9528	0,9475	0,9467
		Promedio		0,9561	0,9521	0,9511

Nota. Evaluación del algoritmo de ensamble bajo condiciones estándar contrastando el uso de unigramas frente a bigramas.

Los resultados de la Tabla 7 muestran que el *random forest*, en su estado base, tiene un rendimiento sólido, pero inferior a la regresión logística, con un F1-Score promedio de 0,9511. En cuanto a la representación, se repite el patrón observado anteriormente: los Unigramas (Exp. 5) superan a los Bigramas (Exp. 6). La caída en el rendimiento al usar N-gramas (0,9467) sugiere que el aumento de dimensionalidad afecta negativamente al modelo de árboles cuando se dispone de un conjunto de datos limitado (1200 registros), lo cual dificulta la generalización.

Finalmente, se aplicó la técnica de *data augmentation* al *random forest* para evaluar su capacidad de mejora. Los resultados se detallan en la Tabla 8.

Tabla 8

Impacto del data augmentation *en el modelo* random forest

N.°	Modelo	Técnica de datos	Representación (TF-IDF)	Precisión	*Recall*	F1-Score
7	*Random forest*	*Data augmentation*	Unigramas (1-gram)	0,9611	0,9583	0,9573
8	*Random forest*	*Data augmentation*	N-gramas (1,2)	0,9615	0,9575	0,9566
		Promedio		0,9613	0,9579	0,9570

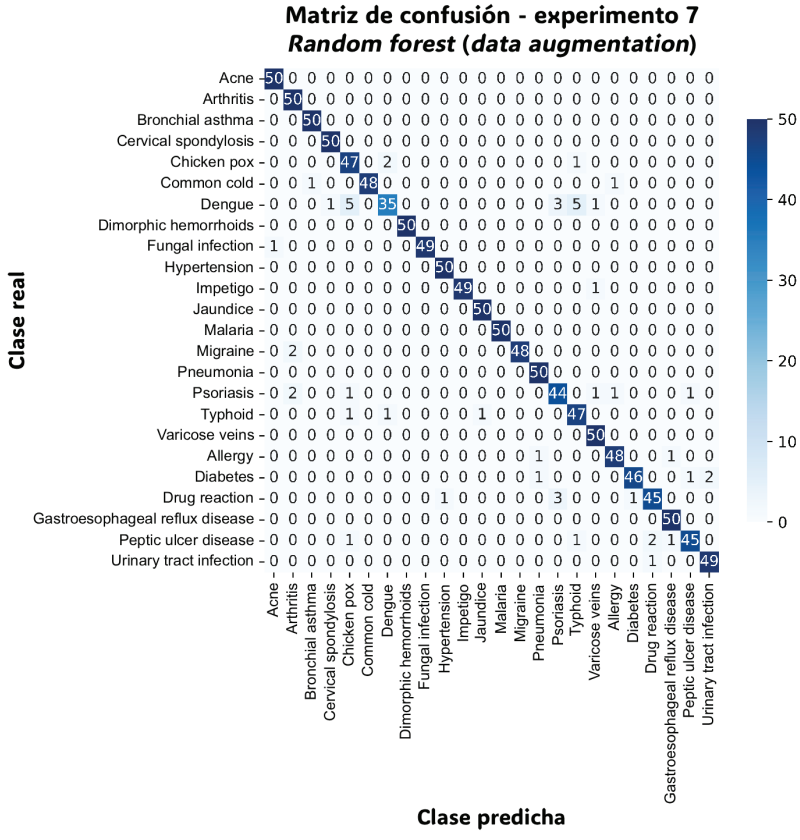
Nota. Métricas de rendimiento obtenidas tras la aplicación de datos sintéticos para mitigar el sobreajuste en el modelo de árboles.

Tal como se aprecia en la Tabla 8, el aumento de datos permitió al *random forest* superar sus métricas base: alcanzó un F1-Score promedio de 0,9570. El mejor escenario para este algoritmo fue el experimento 7 (*data augmentation* + unigramas), con un F1-Score de 0,9573. Aunque la mejora es consistente, la diferencia entre usar unigramas y bigramas se reduce significativamente en presencia de más datos (una diferencia de apenas 0,0007 en F1-Score), lo que indica que el *data augmentation* ayuda a estabilizar el modelo frente a la alta dimensionalidad.

A continuación, se presenta la matriz de confusión del mejor modelo de ensamble obtenido. La comparación entre la matriz de confusión del experimento 3 (regresión logística) y la del experimento 7 (*random forest*) revela un patrón consistente: el modelo de ensamble exhibe mayor dispersión de errores fuera de la diagonal principal, particularmente en enfermedades con vocabulario sintomático compartido. Este comportamiento sugiere que la estructura no lineal del *random forest* tiende a sobreajustar levemente los patrones de clases con pocas instancias cuando el corpus base es reducido (cincuenta registros por clase), incluso con la aplicación de *data augmentation*. Este hallazgo refuerza la recomendación de Pranckevičius & Marcinkevičius (2017) de evaluar sistemáticamente ambos tipos de clasificadores antes de seleccionar el modelo definitivo para aplicaciones clínicas.

Figura 6

Matriz de confusión del mejor modelo random forest *(experimento 7)*



Nota. Distribución de aciertos y confusiones para el algoritmo de ensamble optimizado con unigramas y aumento de datos.

DISCUSIÓN DE LOS RESULTADOS

El análisis comparativo de los ocho escenarios experimentales arroja hallazgos significativos sobre la interacción entre la naturaleza del algoritmo, la técnica de representación y el volumen de datos en el dominio de la clasificación de textos médicos.

En primer lugar, se evidencia una superioridad consistente de la regresión logística sobre el *random forest* para este conjunto de datos específico. El mejor modelo lineal (experimento 3) superó al mejor modelo de ensamble (experimento 7) en más de dos puntos porcentuales en el F1-Score (0,9797 frente a 0,9573). Este hallazgo es coherente con lo expuesto por Genkin et al. (2007), quienes sostienen que los modelos lineales suelen ser altamente efectivos en clasificación de textos, debido a que la alta dimensionalidad de los vectores (TF-IDF) a menudo hace que las clases sean linealmente separables. Mientras que *random forest* destaca por manejar interacciones no lineales complejas (Pal, 2005), en el contexto de descripciones cortas de síntomas, la complejidad del ensamble de árboles parece no aportar una ventaja decisiva frente a la eficiencia de los hiperplanos de decisión de la regresión logística.

Este resultado se alinea con el comportamiento observado en estudios recientes que operan sobre el mismo dominio clínico. Hassan et al. (2024), quienes trabajaron con el mismo corpus Symptom2Disease, reportaron que los modelos de lenguaje profundo como el modelo de normalización de conceptos médicos-representaciones de codificadores bidireccionales de transformadores (*medical concept normalization–bidirectional encoder representations from transformers*, MCN-BERT) superan en exactitud a los clasificadores clásicos; no obstante, los propios autores reconocen que esta ventaja conlleva un costo computacional significativamente mayor y una reducción en la interpretabilidad del modelo, aspecto crítico en entornos de apoyo diagnóstico.

En ese sentido, los resultados del presente estudio indican que, cuando los recursos computacionales son limitados o la transparencia del modelo es un requisito del sistema, la regresión logística representa una alternativa robusta y eficiente que compite favorablemente con arquitecturas más complejas en escenarios de datos pequeños. Esta tesis es consistente con lo reportado por Vyas et al. (2025), quienes encontraron que los modelos lineales bien calibrados con representaciones TF-IDF alcanzan rendimientos competitivos en corpus de tamaño reducido sin sacrificar la capacidad de explicar las predicciones a través de los pesos de sus coeficientes.

En segundo lugar, la aplicación de *data augmentation* demostró ser una estrategia transversalmente efectiva. En ambos algoritmos, el uso de datos sintéticos generados por *random swap* mejoró las métricas de evaluación respecto a la línea base. Esto valida la hipótesis de que, en corpus con un número limitado de muestras por clase (cincuenta registros), la variabilidad léxica artificial actúa como un regularizador potente que ayuda a los modelos a generalizar mejor y evitar el sobreajuste. Este resultado refuerza la

importancia del preprocesamiento robusto en tareas de procesamiento del lenguaje natural (PLN) clínico, en el que la disponibilidad de datos etiquetados suele ser escasa (Báez et al., 2022).

Finalmente, respecto a la representación semántica, los resultados desafiaron la hipótesis inicial de que los N-gramas (bigramas) mejorarían la detección de patologías. Se observó que el uso de unigramas (1-gram) fue marginalmente superior o equivalente al uso de bigramas. Esto sugiere que, en el *dataset* Symptom2Disease, las palabras clave individuales (por ejemplo: *vomiting*, *itching*) poseen una carga discriminativa suficiente por sí mismas. Tal como señalan Aggarwal y Zhai (2012), aumentar la dimensionalidad del espacio de características (agregando bigramas) sin un aumento proporcional en el volumen de datos puede diluir la densidad de información, lo que explica por qué los modelos más simples (unigramas) lograron mantener una precisión tan elevada sin el costo computacional adicional de procesar secuencias de palabras.

En conjunto, los resultados sugieren que para tareas de triaje o diagnóstico preliminar basado en textos breves, un modelo lineal bien calibrado con técnicas de aumento de datos puede ofrecer un rendimiento superior y más eficiente que arquitecturas más complejas. Desde una perspectiva práctica, los resultados de este estudio aportan evidencia empírica relevante para el diseño de sistemas de apoyo al diagnóstico en entornos con recursos limitados. La adopción de un modelo lineal como la regresión logística, combinado con un *pipeline* de preprocesamiento estándar y *data augmentation* mediante *random swap*, ofrece una arquitectura computacionalmente eficiente, interpretable y replicable que facilita la validación por profesionales médicos y su integración en flujos de trabajo clínicos reales.

Asimismo, la alta exactitud alcanzada en las 24 categorías diagnósticas sugiere que este tipo de enfoque podría extenderse a sistemas de triaje automatizado en entornos de telemedicina, en los que la entrada del usuario es una descripción textual libre de sus síntomas (Al-qarni & Algarni, 2025). Sin embargo, la validación de estos resultados en corpus clínicos en español, con mayor variabilidad léxica y solapamiento sintomático, constituye el siguiente paso natural de esta línea de investigación.

CONCLUSIONES

Los resultados obtenidos en este estudio evidencian el potencial y la viabilidad de los modelos clásicos de aprendizaje automático para abordar con eficacia la clasificación automática de textos médicos breves. Tanto la regresión logística como el *random forest* demostraron un desempeño notable, pues superaron el umbral del 95 % de exactitud global tras la aplicación de técnicas de optimización, lo que confirma que, incluso con conjuntos de datos limitados, es posible construir herramientas de apoyo al diagnóstico con alta fiabilidad mediante un preprocesamiento riguroso.

Desde una perspectiva práctica, los hallazgos de este estudio ofrecen una guía metodológica concreta para equipos de investigación que desarrollen herramientas de clasificación de texto médico con recursos limitados. Se recomienda priorizar modelos lineales como la regresión logística cuando el corpus disponible es reducido y el vocabulario diagnóstico es específico; complementar el conjunto de entrenamiento con técnicas de *data augmentation* léxica para mitigar el efecto del bajo número de muestras por clase; y optar por representaciones TF-IDF con unigramas como punto de partida, dado que la incorporación de bigramas puede deteriorar el rendimiento en corpus de pequeñas dimensiones. Estas recomendaciones son especialmente pertinentes en el contexto de la salud digital en economías en desarrollo, donde la escasez de grandes corpus médicos etiquetados obliga a maximizar el rendimiento de los datos disponibles (Calero Sánchez et al., 2024).

En particular, el modelo de regresión logística logró los mejores indicadores de rendimiento de toda la investigación, pues se destacó con un F1-Score de 0,9797 en su configuración óptima. Estos hallazgos confirman su idoneidad para manejar espacios vectoriales de alta dimensionalidad (TF-IDF), en los que la linealidad de sus fronteras de decisión resultó ser más eficaz que las estructuras no lineales de los árboles para discriminar síntomas específicos. Su capacidad para asignar pesos interpretables a términos clave refuerza su valor en aplicaciones biomédicas, en las que la eficiencia computacional es prioritaria.

Por su parte, el algoritmo *random forest* también obtuvo resultados sólidos, alcanzó un F1-Score de 0,9573, aunque mostró una leve inferioridad respecto al modelo lineal. Esta diferencia de rendimiento sugiere que, en el contexto de descripciones cortas y dispersas de síntomas, la complejidad del ensamble de árboles no logra capitalizar sus ventajas habituales sobre la linealidad y puede verse más afectada por la dispersión de los datos. No obstante, su desempeño se mantuvo estable y competitivo validándose como una alternativa robusta frente al ruido.

Un hallazgo transversal y significativo fue el impacto determinante de la técnica de *data augmentation*. El incremento de métricas observado en ambos modelos, al duplicar el corpus mediante intercambio léxico (*random swap*), corrobora que la variabilidad sintética actúa como un regularizador eficaz, pues mitiga el sobreajuste inherente a *datasets* pequeños. Asimismo, el análisis de la representación semántica reveló que los unigramas poseen suficiente carga discriminativa por sí mismos; la incorporación de N-gramas (bigramas) no tradujo mejoras sustanciales, lo que evidencia que la densidad terminológica de los síntomas individuales es el factor predictor dominante en este dominio.

Cabe señalar también un aspecto estructural relevante del corpus analizado: las 24 enfermedades que componen el conjunto de datos presentan cuadros clínicos considerablemente distintos entre sí, de modo que sus síntomas característicos resultan altamente específicos para cada condición. Por ejemplo, las erupciones cutáneas escamosas de la

psoriasis difícilmente se confunden con la fiebre, el dolor articular y el exantema hemorrágico del dengue. Esta especificidad léxica favorece que los modelos alcancen métricas de exactitud elevadas. Sin embargo, en escenarios clínicos reales, donde dos o más enfermedades comparten síntomas comunes (como fiebre, fatiga o dolor muscular), la tarea de clasificación se vuelve considerablemente más compleja y los clasificadores podrían registrar un desempeño inferior.

En consecuencia, la capacidad de generalización de los modelos presentados debe interpretarse en el contexto de un corpus con alta separabilidad interclase. Extender estos resultados a conjuntos de datos con mayor solapamiento sintomático constituye una limitación del presente trabajo y una línea de investigación futura prioritaria, que podría abordarse mediante la incorporación de corpus con mayor complejidad diagnóstica o la exploración de representaciones contextuales de mayor profundidad semántica.

En conjunto, estos resultados respaldan la implementación de flujos de trabajo basados en modelos lineales optimizados y enriquecimiento de datos como estrategia costo-efectiva para la clasificación de literatura médica. A su vez, ponen en evidencia que la clave para optimizar estos sistemas no reside necesariamente en aumentar la complejidad del algoritmo, sino en refinar las estrategias de curación de datos y representación de características sentando una base sólida para futuras investigaciones que busquen integrar estas herramientas en entornos clínicos reales.

REFERENCIAS

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. En C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163-222). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6

Al-qarni, S. S., & Algarni, A. (2025). Disease prediction from symptom descriptions using deep learning and NLP technique. *International Journal of Advanced Computer Science and Applications*, *16*(5), 416-426. https://doi.org/10.14569/IJACSA.2025.0160541

Alsaeedi, A. (2020). A survey of term weighting schemes for text classification. *International Journal of Data Mining, Modelling and Management*, *12*(2), Artículo 237. https://doi.org/10.1504/IJDMMM.2020.10028060

Báez, P., Arancibia, A. P., Chaparro, M. I., Bucarey, T., Núñez, F., & Dunstan, J. (2022). Procesamiento de lenguaje natural para texto clínico en español: el caso de las listas de espera en Chile. *Revista Médica Clínica Las Condes*, *33*(6), 576-582. https://doi.org/10.1016/J.RMCLC.2022.10.002

Barman, N., Karim, F., & Sharma, K. (2023). *Symptom2Disease: Diseases and natural language symptom descriptions*. https://www.kaggle.com/datasets/niyarrbarman/symptom2disease?resource=download

Calero Sánchez, M., González, J. C., Sánchez Berriel, I., Burillo-Putze, G., & Roda García, J. L. (2024). Natural language processing for reviewing search results from PubMed. *Revista Española de Urgencias y Emergencias*, *3*, 184-195. https://doi.org/10.55633/S3ME/REUE030.2024

Cardoso, F. E., Ferneda, E., & Botega, L. (2023). Clasificación de textos: un enfoque con uso de *machine learning*. *Revista EDICIC*, *3*(3), 1-17. https://doi.org/10.62758/re.v3i3.212

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.* DaimlerChrysler. https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic Regression for text categorization. *Technometrics*, *49*(3), 291-304. https://doi.org/10.1198/004017007000000245

Hassan, E., Abd El-Hafeez, T., & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*, *14*(1), Artículo 1507. https://doi.org/10.1038/s41598-024-51615-5

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), Artículo 150. https://doi.org/10.3390/info10040150

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 1-41. https://doi.org/10.1145/3495162

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26* (1), 217-222. https://doi.org/10.1080/01431160412331269698

Powers, D. (2008). Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.*, *2*(1), 37-63. https://doi.org/10.48550/arXiv.2010.16061

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221-232. https://doi.org/10.22364/bjmc.2017.5.2.05

Sun, Y., Li, Y., Qingtao, Z., & Bian, Y. (2020). Application research of text classification based on random forest algorithm. En *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 370-374). IEEE. https://doi.org/10.1109/AEMCSE50948.2020.00086

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), 1-20. https://doi.org/10.1371/journal.pone.0224365

Vyas, D., Shah, M., Kantawala, H., Patel, B., Patel, T., & Enamala, J. (2025). SympTextML: Leveraging natural language symptom descriptions for accurate multi-disease prediction. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(3), 911-924. https://doi.org/10.35882/jeeemi.v7i3.946

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. En K. Inui, J. Jiang, V. Ng & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6382-6388). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1670