

Agente de inteligencia artificial para el apoyo a la primera impresión diagnóstica a partir de descripciones sintomáticas expresadas en lenguaje natural

Artificial Intelligence Agent for Supporting the Initial Diagnostic Impression from Symptom Descriptions Expressed in Natural Language

Juan Pablo Bertel Morales

Tesis de maestría

Asesor, docente

Yomin Estiven Jaramillo Munera

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA
MEDELLÍN
2025

Contenido

1. Introducción	7
2. Planteamiento del problema	8
3. Justificación	10
4.1 Objetivos	12
4.1. General	12
4.2. Específicos	13
5. Marco conceptual.	13
5.1. El diagnóstico médico: naturaleza clínica y epistemológica	13
5.2. Signos y síntomas clínicos: fundamentos y desafíos para la interpretación.	14
5.3. Diagnóstico médico asistido por inteligencia artificial.	15
5.4. Implementaciones de machine learning en el diagnóstico médico.	17
5.5. Consideraciones éticas y privacidad en el uso de IA con datos clínicos .	20
6. Diseño metodológico.	20
6.1. Comprensión del negocio y los datos.	22
6.2. Preparación de los datos y modelación.	23
6.2.1. Modelos de embbings para anamnesis sintéticas.	25
6.2.2. Modelos de clasificación (MLP).	27
7. Resultados.	30
7.1. Resultados del objetivo específico 1: preparación y análisis exploratorio del conjunto de datos.	30
7.1.1. Análisis exploratorio de datos (EDA).	30
7.1.2. Características demográficas de la muestra.	31
7.1.3. Distribución de diagnósticos y evidencias clínicas.	33
7.2. Resultados del objetivo específico 2: desempeño del modelo de clasificación diagnóstica.	36
7.3. Resultados del objetivo específico 3: prueba de estrés y evaluación del agente conversacional.	39
8. Conclusiones.	44
9. Referencias	
10. anexos	49

Agradecimientos

En primer lugar, quiero expresar mi profundo agradecimiento a mis padres, Jairo de Jesús Bertel Beltrán y Adela Morales Chiquillo, por los valores, el apoyo incondicional y la confianza que han guiado mi camino académico y personal. Extiendo igualmente mi gratitud a mi hermano, Miguel Ángel Romero Puello, y a mis abuelas y abuelos, cuyo cariño y acompañamiento constante han sido un apoyo esencial y una fuente permanente de motivación.

Agradezco a mis profesores Yomin Estiven Jaramillo y Juan David Martinez por sus valiosas asesorías y orientación a lo largo de este proceso, así como a todos los docentes que han contribuido a mi formación académica y a mi comprensión de la ciencia de datos.

De manera especial, agradezco a San Vicente Fundación por el apoyo brindado durante mi formación de posgrado, en particular, a Sebastián Correal Franco, uno de mis jefes, por impulsarme a emprender este proyecto y respaldar su desarrollo, y al doctor Juan Gonzalo Fernandez, cuyas observaciones desde la práctica clínica fueron clave para el diseño de la interfaz y del modelo.

Por último, pero no menos importante, agradezco a mis gatos, Tommy y Leo, cuya compañía silenciosa pero constante me acompañó en todas y cada una de las noches de trabajo. Su presencia hizo más llevadero el esfuerzo y me recordó, incluso en los momentos de mayor cansancio, la importancia de los pequeños apoyos en los grandes procesos.

Lista de figuras

Figura 1. Etapas del proceso CRISP-DM aplicadas al desarrollo del sistema de diagnóstico asistido por IA.	21
Figura 2. Codificación de síntomas y generación de anamnesis sintética.	24
Figura 3. Arquitectura del perceptrón.	27
Figura 4. Distribución de edades en la población de pacientes.	31
Figura 5. Distribución por sexo.	32
Figura 6. Distribución de edad por sexo.	33
Figura 7. Top 10 diagnósticos más frecuentes.	34
Figura 8. Distribución de sexo por diagnostico (Top 10).	35
Figura 9. Distribución del número de síntomas reportados por paciente.	36
Figura 10. Precisión Top-1 por modelo (marcadores circulares).	37
Figura 11. Interfaz inicial.	40
Figura 12. Cargue modelos.	41
Figura 13. Resultados 1.	42
Figura 14. Resultados 2.	42

Lista de tablas

Tabla 1. Descripción de variables.	23
Tabla 2. Desempeño de modelos AARLC en DDxPlus	39

Resumen

Esta tesis propone el desarrollo de un agente de inteligencia artificial (IA) capaz de apoyar la generación de una primera impresión diagnóstica a partir de síntomas expresados en lenguaje natural. El proyecto parte del reconocimiento de que el diagnóstico médico es una tarea compleja y propensa a errores, especialmente cuando depende de descripciones subjetivas y no estructuradas.

Con el fin de apoyar la toma de decisiones clínicas, se aplicaron técnicas de procesamiento de lenguaje natural y aprendizaje automático bajo la metodología CRISP-DM. El modelo se entrenó utilizando el conjunto de datos sintéticos DDxPlus, lo que permitió simular escenarios clínicos sin comprometer información real de pacientes.

En el proceso, los síntomas fueron transformados en anamnesis sintéticas mediante normalización semántica y posteriormente vectorizados con diferentes modelos de embeddings biomédicos. Estas representaciones se emplearon para entrenar un modelo supervisado encargado de asociar cada narrativa con el diagnóstico confirmado.

Como evaluación adicional, se llevó a cabo una “prueba de estrés” en un entorno simulado, en la cual un profesional de la salud interactuó directamente con el sistema para valorar su capacidad de interpretar descripciones sintomáticas reales y generar sugerencias diagnósticas preliminares de manera coherente, consistente y segura.

Palabras clave: Primera impresión diagnóstica, procesamiento del lenguaje natural, embeddings, clasificación supervisada, diagnóstico asistido por IA.

Abstract

This thesis proposes the development of an artificial intelligence (AI) agent capable of supporting the generation of an initial diagnostic impression based on symptoms expressed in natural language. The project is grounded in the recognition that medical diagnosis is a complex task prone to errors, particularly when it relies on subjective and unstructured descriptions.

To support clinical decision-making, natural language processing and machine learning techniques were applied following the CRISP-DM methodology. The model was trained using the synthetic DDxPlus dataset, which enabled the simulation of clinical scenarios without compromising real patient information.

In the process, symptoms were transformed into synthetic anamneses through semantic normalization and subsequently vectorized using various biomedical embedding models. These representations were then used to train a supervised model tasked with associating each narrative with the confirmed diagnosis.

As an additional evaluation, a “stress test” was conducted in a simulated environment, in which a healthcare professional interacted directly with the system to assess its ability to interpret real symptomatic descriptions and generate preliminary diagnostic suggestions in a coherent, consistent, and safe.

Keywords:Initial diagnostic impression, natural language processing, embeddings, supervised classification, AI-assisted diagnosis.

1. INTRODUCCIÓN

El diagnóstico médico es una de las tareas principales en la atención en salud, ya que requiere integrar múltiples elementos clínicos y de laboratorios, criterios probabilísticos e interpretaciones subjetivas, todo ello en contextos de alta incertidumbre y presión de tiempo. A pesar de los avances en guías clínicas, sigue siendo una etapa susceptible a errores que pueden afectar la calidad del tratamiento y los resultados en el paciente.

Un reto particular en este proceso es la interpretación de los síntomas reportados por los pacientes, entendidos no como valoraciones objetivas, sino como experiencias subjetivas que adquieren significado según la percepción y el contexto de cada individuo (Rosendal et al., 2013). Como señalan estos autores, la manera en que un paciente describe sus molestias está influenciada por factores personales, emocionales y socioculturales, lo que genera una notable variabilidad en el lenguaje empleado para expresar sensaciones como dolor, presión o malestar. Esta diversidad narrativa podría dificultar su estandarización y valoración sistemática en la práctica clínica, especialmente en entornos de atención primaria o con recursos limitados, donde los profesionales deben interpretar con relativa rapidez relatos sintomáticos heterogéneos y altamente contextualizados.

Ante esta problemática, la inteligencia artificial puede ofrecer un apoyo complementario a la toma de decisiones clínicas mediante modelos capaces de identificar patrones en los síntomas y generar posibles primeras impresiones diagnósticas. La evidencia reciente muestra que los sistemas de IA pueden aproximarse al desempeño diagnóstico del personal de salud en diversas tareas, pero que su utilidad radica precisamente en servir como herramienta de apoyo y no como sustituto del juicio médico (Jeong et al., 2025). Según este artículo, los modelos de IA pueden ampliar la cobertura diagnóstica, mejorar la consistencia y reducir la variabilidad, siempre que su

uso se integre dentro de un enfoque de colaboración humano–máquina supervisada por profesionales de la salud.

Esta tesis tiene como propósito desarrollar un agente de IA capaz de generar una primera impresión diagnóstica a partir de síntomas expresados en lenguaje natural por los pacientes. Para ello, se emplean técnicas de normalización semántica y aprendizaje automático que permiten transformar las descripciones sintomáticas en representaciones estructuradas y útiles para el modelado clasificatorio. Siguiendo los lineamientos de la metodología CRISP-DM, el sistema se adapta al contexto clínico con el fin de sistematizar la interpretación de síntomas subjetivos, reduciendo la ambigüedad inherente al lenguaje natural y contribuyendo a mejorar la precisión, consistencia y eficiencia del proceso diagnóstico.

2. PLANTEAMIENTO DEL PROBLEMA

El diagnóstico médico es un proceso complejo que depende en gran medida de la interpretación de síntomas reportados por los pacientes. Estos síntomas, al ser subjetivos, variables y frecuentemente ambiguos, presentan desafíos importantes para su sistematización y análisis cuantitativo. Si bien las tecnologías basadas en IA han demostrado un rendimiento sobresaliente en tareas como la interpretación de imágenes médicas, donde los modelos de deep learning alcanzan áreas bajo la curva (AUC) superiores al 90% (Aggarwal et al., 2021), y en la predicción de desenlaces clínicos a partir de datos estructurados mediante modelos de aprendizaje automático (Shamout et al., 2021), su aplicación en entornos clínicos y, en particular, en la comprensión directa de síntomas expresados en lenguaje natural por el paciente continúa siendo un campo aún por desarrollar. Como advierte Varghese (2020) la adopción clínica a gran escala de sistemas de IA sigue enfrentando barreras importantes, entre ellas la necesidad de validaciones clínicas rigurosas, la interpretabilidad de los modelos, las dificultades de

interoperabilidad y la disponibilidad de datos estructurados y de evidencia robusta que respalde su uso seguro y efectivo.

En los últimos años han surgido múltiples modelos basados en técnicas de aprendizaje automático y procesamiento de lenguaje natural capaces de analizar información clínica en texto libre. Estos desarrollos han permitido, por ejemplo, la extracción automatizada de síntomas, signos y antecedentes desde historias clínicas electrónicas en inglés (Wu et al., 2022), así como el análisis de narrativas médicas escritas en español dentro del sistema público de salud en Chile (Báez et al., 2022). Sin embargo, su adopción real en entornos asistenciales aún enfrenta barreras significativas. Esta limitación no se debe a la ausencia de avances técnicos, sino a obstáculos metodológicos y operativos que impiden su integración efectiva en la práctica clínica. Entre estos podemos observar, la escasez de validaciones externas, el uso de conjuntos de datos homogéneos en las fases de entrenamiento, la ausencia de estudios multicéntricos, la desalineación con los flujos de trabajo clínicos y preocupaciones relacionadas con la transparencia y la privacidad. Como resultado, persiste una brecha considerable entre el rendimiento demostrado por la inteligencia artificial en entornos controlados y su efectividad en escenarios clínicos reales. El Arab et al. (2025) advierten que cerrar esta brecha requiere enfoques interdisciplinarios, estudios observacionales y ensayos pragmáticos que evalúen estas tecnologías en condiciones sanitarias diversas y concretas.

Además de los retos técnicos, el desarrollo de sistemas capaces de interpretar síntomas en lenguaje natural y generar sugerencias para una primera impresión diagnóstica implica importantes consideraciones éticas. Es indispensable garantizar la protección de la privacidad de los datos clínicos, asegurar la transparencia en el funcionamiento del modelo y evitar que estas tecnologías sean utilizadas como sustitutos del juicio clínico. En este sentido, cualquier

aproximación basada en IA debe concebirse como un mecanismo de apoyo para los profesionales de la salud, no como una herramienta de diagnóstico autónomo.

En este contexto, se identifica la necesidad de diseñar y evaluar un agente de IA que contribuya al análisis preliminar de síntomas expresados en lenguaje natural y apoye en la generación de una primera impresión diagnóstica. Una herramienta de este tipo podría fortalecer el proceso de atención médica, agilizar la toma inicial de decisiones y aportar coherencia en la interpretación de descripciones sintomáticas, siempre enfocado desde una perspectiva técnicamente viable y éticamente responsable.

3. JUSTIFICACIÓN

Los errores diagnósticos representan una fuente importante de daño evitable en la atención médica, con efectos sobre la seguridad del paciente, los costos del sistema y la calidad del cuidado clínico. De acuerdo con la Agency for Healthcare Research and Quality (AHRQ), entre el 6 % y el 17 % de los eventos adversos hospitalarios están asociados a fallas en el diagnóstico, siendo una proporción significativa de ellos prevenible (AHRQ, 2020). Estos errores no solo podrían afectar la seguridad del paciente, sino que también incrementarían los costos asistenciales, prolongarían hospitalizaciones innecesarias y deteriorarían la confianza en el sistema de salud.

Por tanto, los errores diagnósticos constituyen una preocupación central en la práctica clínica debido a su impacto directo en la evolución y los desenlaces de los pacientes. Estos errores, además de ser difíciles de detectar y prevenir, pueden comprometer gravemente la seguridad del paciente al conducir a tratamientos inapropiados, demoras terapéuticas o complicaciones evitables. Según un análisis reciente publicado en BMJ Quality & Safety, se estima que alrededor

de 795.000 personas al año sufren daño grave en Estados Unidos como resultado de un error diagnóstico, incluyendo aproximadamente 371.000 muertes y cientos de miles de casos de discapacidad permanente (Newman-Toker et al, 2023).

Una de las principales fuentes de error diagnóstico reside en la subjetividad inherente a los síntomas reportados por los pacientes, particularmente aquellos que no pueden medirse de forma objetiva, como el dolor, la fatiga, el mareo o la sensación febril. Estos síntomas dependen del relato individual, el contexto cultural y la forma en que se expresan, lo que dificulta su cuantificación, estandarización y análisis sistemático. Esta variabilidad introduce ambigüedad clínica y aumenta la probabilidad de imprecisiones en el diagnóstico, especialmente en las etapas iniciales del proceso asistencial. Tal como lo señalan diversos estudios, esta complejidad es especialmente notoria en entornos ambulatorios y de atención primaria, donde las decisiones diagnósticas deben tomarse con información clínica limitada y bajo presión de tiempo (Graber, 2013; Singh & Weingart, 2009). En el contexto colombiano, esta problemática puede verse acentuada por factores estructurales del sistema de salud, como la alta demanda asistencial, la escasez de especialistas y las barreras de acceso geográfico y social a servicios diagnósticos oportunos.

En este contexto, la IA representa una herramienta prometedora para apoyar el razonamiento clínico, particularmente en la normalización semántica de síntomas subjetivos y en la sugerencia de diagnósticos probables a partir de patrones previamente aprendidos. Modelos basados en aprendizaje automático permiten interpretar variaciones en la forma de describir síntomas y asociarlas con diagnósticos clínicos frecuentes, mejorando así la eficiencia del proceso diagnóstico.

Si bien la inteligencia artificial ofrece un gran potencial para apoyar el razonamiento clínico, su implementación en contextos sanitarios exige una evaluación rigurosa no solo de su rendimiento técnico, sino también de su viabilidad ética y práctica. La literatura reciente enfatiza que el éxito de estas tecnologías depende tanto de su desempeño como de dimensiones como la transparencia, la confianza, la protección de la privacidad, la equidad y la calidad de los datos (Jeyaraman et al., 2023). En este sentido, la explicabilidad de las decisiones generadas por los sistemas de IA y la gestión adecuada de riesgos como los sesgos algorítmicos, la seguridad de la información y las brechas de confianza resultan determinantes para su adopción efectiva en la práctica clínica.

En vista de lo anterior, esta tesis se justifica por la necesidad de diseñar y evaluar un modelo de IA que, a partir de los síntomas que experimente y comunique un paciente, sugiera una primera impresión diagnóstica con un enfoque de apoyo a la decisión clínica. Esta herramienta busca complementar, no sustituir, el juicio médico, y contribuir a una atención más equitativa, precisa y eficiente, especialmente en escenarios de alta presión asistencial o acceso limitado a especialistas.

4.1 OBJETIVOS

4.1. GENERAL

Desarrollar un agente basado en inteligencia artificial que, a partir de la descripción de síntomas expresados en lenguaje natural por el paciente, sugiera una primera impresión diagnóstica, con el propósito de apoyar a los profesionales de la salud en la toma de decisiones clínicas iniciales.

4.2. ESPECÍFICOS

1. Preparar el conjunto de datos clínicos para su uso en modelos de aprendizaje automático, mediante el análisis exploratorio de su estructura, la normalización semántica de síntomas, la codificación de diagnósticos y la partición en subconjuntos de entrenamiento y prueba.
2. Construir un modelo de IA que integre técnicas de procesamiento de lenguaje natural y aprendizaje supervisado, capaz de generar una primera impresión diagnóstica a partir de descripciones sintomáticas expresadas en lenguaje natural.
3. Evaluar el desempeño del modelo propuesto mediante métricas cuantitativas y una prueba de estrés en un entorno simulado, incorporando el sistema en un agente conversacional interactivo capaz de recibir entradas en lenguaje natural, interpretar síntomas y ofrecer posibles diagnósticos preliminares con base en evidencia clínica, con el fin de valorar su utilidad como herramienta de apoyo en la toma de decisiones durante la primera impresión diagnóstica.

5. MARCO CONCEPTUAL.

5.1. EL DIAGNÓSTICO MÉDICO: NATURALEZA CLÍNICA Y EPISTEMOLÓGICA

El diagnóstico médico es el acto mediante el cual un profesional de la salud identifica una enfermedad o condición clínica en un paciente a partir de una evaluación racional de sus signos, síntomas, historia clínica y resultados de otras pruebas complementarias. Se trata de una de las funciones centrales de la práctica médica, en la medida en que orienta la selección de tratamientos apropiados, contribuye a la prevención de complicaciones y condiciona los desenlaces clínicos.

Además, Díez y de la Coteria (2010) señalan que el diagnóstico no solo es un proceso técnico o clínico, sino también un problema epistemológico que implica interpretar fenómenos clínicos dentro de un marco científico, social y ético. El acto diagnóstico exige integrar evidencia incompleta, gestionar incertidumbre y justificar las decisiones adoptadas; por tanto, constituye un escenario en el que herramientas computacionales de apoyo pueden contribuir a fortalecer la consistencia y transparencia del razonamiento clínico, sin sustituir el juicio profesional.

5.2. SIGNOS Y SÍNTOMAS CLÍNICOS: FUNDAMENTOS Y DESAFÍOS PARA LA INTERPRETACIÓN.

En el proceso diagnóstico es fundamental distinguir entre signos y síntomas clínicos, dado que representan fuentes complementarias de información. Los signos clínicos son manifestaciones objetivas de una enfermedad que pueden ser observadas, medidas o verificadas por el personal de salud mediante el examen físico, pruebas de laboratorio o técnicas de imagen. Entre ellos se incluyen, por ejemplo, la fiebre, el enrojecimiento, alteraciones en los niveles de glucosa en sangre o lesiones visibles en estudios radiológicos. En el caso del síndrome de ojo seco, Barlett et al. (2015) señalan que algunos de los signos comúnmente evaluados en los estudios incluyen el tiempo de ruptura lagrimal, la tinción con fluoresceína y la osmolaridad lagrimal.

A diferencia de los signos, los síntomas, como se definió anteriormente, corresponden a manifestaciones subjetivas que el paciente experimenta y comunica al profesional, pero que no son directamente observables. En el mismo estudio, los autores describen síntomas como ardor, escozor, visión borrosa y molestia ocular, cuya intensidad y presencia pueden variar ampliamente entre pacientes, incluso cuando los signos clínicos son similares o mínimos. Esta disociación entre lo que el paciente percibe y lo que puede medirse objetivamente define el carácter subjetivo del síntoma y constituye uno de los principales desafíos para el diagnóstico médico y, en particular, para su automatización. Los modelos computacionales destinados a apoyar el

diagnóstico deben ser capaces de manejar esta variabilidad expresiva y semántica, lo cual resulta especialmente relevante en proyectos, como el presente, que se centran en el análisis de descripciones sintomáticas en lenguaje natural.

Si bien los signos clínicos proporcionan datos objetivos, el artículo antes mencionado demuestra que presentan una baja correlación con los síntomas del paciente, lo que limita su utilidad como predictores únicos del diagnóstico. Esta observación confirma la necesidad de métodos que incorporen datos clínicos observables y sensaciones subjetivas, dada la complejidad intrínseca del proceso diagnóstico. Esta riqueza, derivada de la tensión entre lo medible y lo percibido, abre la posibilidad de incorporar nuevos recursos computacionales que faciliten la toma de decisiones clínicas mediante el análisis de datos. En este sentido, el aprendizaje automático y la IA se presentan como tecnologías muy prometedoras para habilitar modelos capaces de reconocer patrones complejos y realizar predicciones diagnósticas.

5.3. DIAGNÓSTICO MÉDICO ASISTIDO POR INTELIGENCIA ARTIFICIAL.

Según Kaur et al. (2022), el diagnóstico médico asistido por inteligencia artificial puede entenderse como el uso de técnicas predictivas de IA para automatizar la identificación de enfermedades a partir de grandes volúmenes de datos clínicos, datos de laboratorio, imágenes médicas u otras fuentes del paciente, con el objetivo de apoyar la toma de decisiones clínicas y mejorar la precisión, velocidad y eficiencia del proceso diagnóstico. Históricamente, esta tarea ha requerido la combinación de diferentes fuentes de información y el criterio médico basado en la experiencia. Con la creciente cantidad de datos disponibles, se dificultará el analizar toda la información relevante en su totalidad. Los algoritmos de aprendizaje automático se presentan aquí como herramientas muy útiles, capaces de automatizar y optimizar la interpretación de datos clínicos estructurados y no estructurados.

En los últimos años, la IA ha apoyado diversos campos del conocimiento, incluyendo la atención médica, en particular en el apoyo al diagnóstico médico. Métodos como el aprendizaje automático han contribuido al desarrollo de modelos predictivos capaces de manipular grandes cantidades de datos clínicos con gran precisión. Hay indicios claros de que la IA desempeñará un papel fundamental en la mejora del proceso diagnóstico en aquellas áreas de la patología donde la interpretación visual o el análisis de patrones complejos resultan críticos.

Diversos estudios han documentado aplicaciones exitosas de estas tecnologías en contextos clínicos reales. Un ejemplo relevante, es el estudio de Wu et al. (2022), quienes realizaron una evaluación sobre la aplicación de la IA en el diagnóstico y tratamiento clínico. El análisis incluyó 40 revisiones que, en conjunto, evaluaron más de 1.000 estudios primarios en 31 enfermedades distintas. Los autores reportaron que los modelos de IA lograron una tasa de aciertos superior al 90% en la detección de condiciones como úlceras gastrointestinales, hemorragias digestivas, tumores esofágicos, gástricos e intestinales. Estos resultados muestran el potencial de la IA como herramienta de apoyo clínico, con la capacidad de simplificar el proceso de diagnóstico, reducir los márgenes de error y generar una atención más rápida y eficiente de cara al paciente.

Sin embargo, diversos estudios han señalado que los sistemas de IA, como los modelos de lenguaje de gran escala, no deben concebirse como reemplazos del juicio clínico humano, sino como herramientas de apoyo complementario. Por ejemplo, D'Souza et al. (2023) evaluaron el desempeño de ChatGPT 3.5 en 100 casos clínicos de psiquiatría, encontrando que en el 61 % de los casos obtuvo una calificación de desempeño excelente (Diagnósticos plausibles, propuestas terapéuticas juicio compatible con buenas prácticas y omisión de recomendaciones poco éticas). Estos hallazgos refuerzan la idea de que la IA puede servir como un asistente cognitivo útil para apoyar el razonamiento clínico.

Ahora bien, el uso de IA y modelos de machine learning en la predicción de diagnósticos médicos implica necesariamente el procesamiento de grandes volúmenes de datos sensibles, como historiales clínicos, síntomas reportados y antecedentes personales. Esto plantea importantes desafíos en términos de privacidad, seguridad y confidencialidad. Jeyaraman et al. (2023) destacan que, en un entorno sanitario impulsado por datos, proteger la privacidad del paciente es esencial, ya que la exposición indebida de información médica puede generar consecuencias psicológicas adversas y socavar la confianza en los sistemas de salud. Además, subrayan la importancia de implementar soluciones como el cifrado homomórfico y la computación multipartita segura, que permiten procesar datos sin comprometer su confidencialidad. En este sentido, la incorporación de modelos predictivos debe ir acompañada de marcos de gobernanza sólidos que aseguren el cumplimiento normativo, la protección de la información personal y el respeto por los principios éticos en la atención médica.

5.4. IMPLEMENTACIONES DE MACHINE LEARNING EN EL DIAGNÓSTICO MÉDICO.

Una demostración concreta del potencial del machine learning en la práctica clínica fue presentada por J. A. M. Sidey-Gibbons y C. J. Sidey-Gibbons (2019), quienes desarrollaron y compararon distintos algoritmos de clasificación aplicados al diagnóstico de cáncer de mama. Utilizando un conjunto de datos públicos relacionados a los pacientes de cáncer de mama, entrenaron tres modelos supervisados: regresión logística regularizada con LASSO, máquinas de soporte vectorial, y redes neuronales artificiales, todos ellos basados en características morfológicas de núcleos celulares obtenidas mediante aspiración con aguja fina.

Los modelos fueron evaluados sobre una muestra de validación independiente, alcanzando niveles de precisión entre el 94 % y el 96 %. El algoritmo basado en máquinas de soporte vectorial fue el que mostró mayor rendimiento (exactitud = 0.96; AUC = 0.97), y su desempeño

mejoró marginalmente al combinarse con los otros dos modelos mediante un esquema de votación (ensemble), logrando una precisión total del 97 % y una especificidad del 95 %. Este ejemplo evidencia cómo algoritmos de aprendizaje automático, incluso con arquitecturas relativamente simples, pueden ofrecer soporte diagnóstico automatizado con niveles de desempeño altos. Estos resultados no solo destacan la efectividad de los modelos predictivos, sino también su aplicabilidad en tareas clínicas reales, siempre que se cuente con datos estructurados de calidad y se implemente un proceso riguroso de entrenamiento, validación y evaluación.

Más allá del diagnóstico basado en características celulares, la IA también ha mostrado un impacto significativo en otros campos clínicos, como el análisis genético y fenotípico. Una aplicación notable de esto se encuentra en el ámbito del diagnóstico genómico. En su revisión, Dias y Torkamani (2019) documentan múltiples casos en los que modelos de aprendizaje profundo, particularmente redes neuronales convolucionales, han logrado rivalizar con las capacidades humanas en tareas de diagnóstico a partir de imágenes médicas. Por ejemplo, describen el uso de DeepGestalt, un sistema basado en redes neuronales convolucionales capaz de identificar síndromes genéticos raros a partir del análisis automatizado de fotografías faciales. Este tipo de herramientas ha sido integrado a plataformas como PEDÍA, que combinan datos genéticos y fenotípicos para priorizar variantes patogénicas con una precisión comparable a la de expertos humanos, alcanzando un 97% de recall y un 99% de precisión. Este caso ejemplifica el potencial de la IA y el aprendizaje automático no sólo en tareas visuales, sino en el cruce de múltiples modalidades de datos clínicos para apoyar diagnósticos de manera precisa y rápida.

En conclusión, el diagnóstico médico es una actividad compleja que consiste en integrar el análisis de signos clínicos objetivos con síntomas subjetivos dentro de un marco de

interpretación científica. Esta dualidad entre lo observable y lo percibido por el paciente constituye un desafío intrínseco para la práctica médica, especialmente cuando los signos no se correlacionan con la sintomatología (Barlett et al., 2015). Frente a esta limitación, la necesidad de herramientas que integren de forma eficiente múltiples tipos de datos ha dado paso al interés por soluciones tecnológicas que refuercen la precisión y consistencia del proceso diagnóstico.

En este contexto, la IA y el aprendizaje automático, han mostrado un enorme potencial como herramienta de apoyo clínico. Su capacidad para analizar grandes volúmenes de datos estructurados y no estructurados ha permitido el desarrollo de modelos predictivos capaces de detectar enfermedades con altos niveles de precisión. Tal como se evidencia en el estudio de Wu et al. (2022). Sin embargo, diversos autores coinciden en que la IA debe entenderse como un asistente cognitivo y no como un reemplazo del juicio clínico humano (D'Souza et al., 2023).

Aplicaciones clínicas concretas, como los modelos empleados en el diagnóstico de cáncer de mama (Sidey-Gibbons & Sidey-Gibbons, 2019), o los sistemas de reconocimiento facial aplicados al diagnóstico de síndromes genéticos (Dias & Torkamani, 2019), demuestran que los algoritmos pueden alcanzar niveles de desempeño altos cuando se dispone de datos clínicos de calidad. Estos ejemplos ilustran el valor de integrar múltiples fuentes de información (morfológica, fenotípica y genómica) en escenarios clínicos con enfoques automatizados. Sin embargo, estos sistemas deben implementarse de forma progresiva, validada y contextualizada para determinar su verdadera utilidad clínica y su adopción por parte del personal médico.

5.5. CONSIDERACIONES ÉTICAS Y PRIVACIDAD EN EL USO DE IA CON DATOS CLÍNICOS .

Finalmente, la aplicación de la IA al diagnóstico médico también conlleva algunos desafíos éticos, como, el uso de datos clínicos sensibles respaldados por marcos de gobernanza sólidos que garanticen la privacidad, el consentimiento informado y la autonomía del paciente. Como

advierten Jeyaraman et al. (2023), la implementación de estas tecnologías debe estar acompañada de políticas rigurosas de seguridad y confidencialidad, dado que un manejo inadecuado de la información médica puede comprometer la privacidad del paciente, generar consecuencias psicológicas negativas y debilitar la confianza en el sistema de salud. Solo bajo el respaldo de marcos éticos sólidos y soluciones tecnológicas orientadas a la protección de datos, la IA podrá consolidarse como una herramienta confiable, segura y éticamente aceptable dentro del ecosistema médico contemporáneo.

6. DISEÑO METODOLÓGICO.

El presente estudio adopta como base la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente reconocida por su utilidad en la estructuración de proyectos de ciencia de datos. A más de dos décadas de su creación, CRISP-DM se mantiene como un marco metodológico considerablemente empleado en la industria y la investigación, debido a su flexibilidad, enfoque orientado a procesos y capacidad de adaptación a diferentes dominios aplicados (Martinez-Plumed et al., 2019; Ayele et al., 2020). Estas características lo convierten en una guía adecuada para abordar retos como el que plantea este escrito: el diseño de un agente de inteligencia artificial que apoye la formulación de la primera impresión diagnóstica a partir de la descripción de síntomas en lenguaje natural.

En este proyecto, la metodología CRISP-DM es adaptada al contexto del campo de la salud, para orientar el desarrollo del agente propuesto, especialmente en escenarios con alta demanda asistencial o acceso limitado a personal sanitario especializado. La Figura 1 ilustra las seis fases del proceso metodológico, contextualizadas para el desarrollo de este proyecto y organizadas en

un flujo iterativo que refleja la naturaleza cíclica del aprendizaje automático aplicado a problemas clínicos.

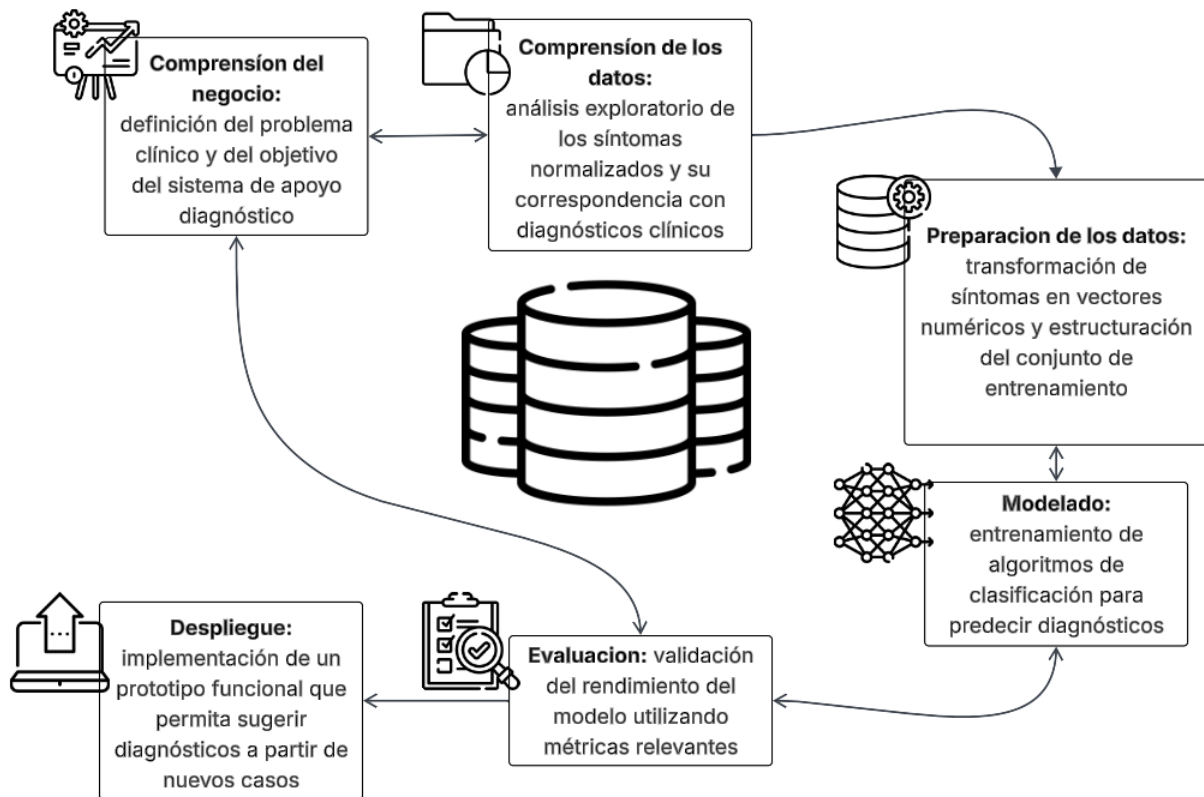


Figura 1. Etapas del proceso CRISP-DM aplicadas al desarrollo del sistema de diagnóstico asistido por IA.
Nota. Fuente: Elaboración propia

6.1. COMPRESIÓN DEL NEGOCIO Y LOS DATOS.

Las dos primeras fases de la metodología CRISP-DM (comprensión del negocio y comprensión de los datos) constituyen el punto de partida metodológico para el cumplimiento del primer objetivo específico; preparar el conjunto de datos clínicos para su uso en modelos de aprendizaje. En la fase de comprensión del negocio se plantea el problema clínico central del proyecto, el cual gira en torno a la dificultad inherente a la interpretación de síntomas subjetivos expresados en lenguaje natural por los pacientes, debido a su ambigüedad, variabilidad lingüística y cultural y

escasa estructuración, lo cual puede ser un limitante en su aprovechamiento en sistemas computacionales.

En la fase de comprensión de los datos se realizó un análisis exploratorio del conjunto DDXPlus, una base de datos sintética de libre acceso construida para apoyar el desarrollo de sistemas automáticos de diagnóstico médico. Este conjunto fue desarrollado por el Mila-Quebec AI Institute en colaboración con Dialogue Health Technologies Inc., y presentado en la NeurIPS 2022 como parte del track de Datasets and Benchmarks. El dataset contiene aproximadamente 1.3 millones de registros sintéticos de pacientes, generados a partir de una base de conocimiento médico propietaria construida sobre más de 20,000 publicaciones clínicas, y revisada por expertos médicos. Cada registro simula un caso clínico realista incluyendo variables sociodemográficas, síntomas (binarios, categóricos y multiselección), antecedentes, diagnóstico confirmado y diagnóstico diferencial asociado (Fansi Tchango et al., 2022).

La generación de los pacientes se basó en una combinación de datos censales de EE.UU. (New York Census 2010–2015), modelos probabilísticos condicionados por edad, sexo y región geográfica, y un sistema de diagnóstico automatizado basado en reglas clínicas. Además, el sistema garantiza que los diagnósticos diferenciales contengan al menos la patología subyacente del paciente, lo cual permite una validación más robusta y realista del desempeño de modelos de clasificación en tareas de diagnóstico automatizado. Por estas características, DDXPlus constituye una fuente altamente estructurada y representativa para entrenar, validar y probar agentes en escenarios clínicos simulados.

Los datos se encuentran estructurados en variables discretas y categóricas, junto con atributos de texto estandarizado para describir evidencias clínicas. La siguiente tabla resume las principales variables incluidas en el conjunto de datos:

Variable	Descripción
“AGE”	Edad del paciente en años
“SEX”	Sexo biológico del paciente; M: Masculino, F: Femenino
“PATHOLOGY”	Diagnóstico confirmado o enfermedad principal del paciente
“DIFFERENTIAL_DIAGNOSIS”	Lista ordenada de diagnósticos diferenciales con sus respectivas probabilidades de ocurrencia.
“EVIDENCES”	Lista de síntomas y antecedentes experimentados y reportados por el paciente
“INITIAL_EVIDENCE”	Primer síntoma reportado por el paciente.

Tabla 1. Descripción de variables

6.2. PREPARACIÓN DE LOS DATOS Y MODELACIÓN.

Las etapas de preparación de datos y modelado constituyen el núcleo técnico del desarrollo del agente, en línea con el segundo objetivo específico de este escrito: construir un modelo de IA que integre técnicas de procesamiento de lenguaje natural y aprendizaje supervisado para generar una primera impresión diagnóstica a partir de descripciones sintomáticas expresadas en lenguaje natural.

La fase de preparación se enfocó en convertir las evidencias clínicas estructuradas del dataset DDxPlus en entradas textuales aptas para un modelo de clasificación que busca relacionar el diagnóstico confirmado del paciente con sus síntomas. Cada registro de paciente en DDxPlus incluye una lista de evidencias (EVIDENCES), representadas por códigos alfanuméricos que identifican síntomas y antecedentes en formato estandarizado (por ejemplo: E_48, E_54_@_V_161). Para alimentar adecuadamente un modelo de Natural Language Processing

(NLP), fue necesario traducir dichos códigos simbólicos a frases clínicas canónicas legibles, construyendo así un campo de texto sintético (note_text) que compendia los síntomas reportados en lenguaje natural.

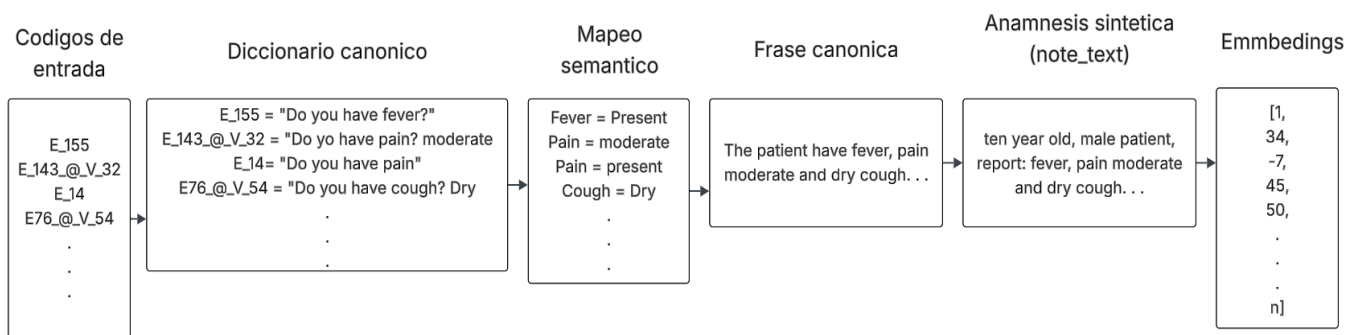


Figura 2. Codificación de síntomas y generación de anamnesis sintética.
 Nota. Fuente: Elaboración propia

El proceso de transformación de las evidencias clínicas en un texto interpretable se resume en la Figura 2, donde se muestra el flujo completo desde la estructura original del dataset hasta la generación de la anamnesis sintética final. En primer lugar, cada código presente en la variable EVIDENCES fue interpretado mediante el diccionario canónico proporcionado por DDxPlus, el cual asigna a cada evidencia una frase clínica estandarizada o una pregunta asociada a la presencia de un síntoma en concreto. Posteriormente, durante la etapa de mapeo semántico, estas frases se descompusieron en atributos clínicos normalizados (por ejemplo: fever = present, pain = moderate, cough = dry), lo que permitió representar los síntomas de manera consistente y libre de ambigüedades.

Una vez obtenidos los atributos clínicos normalizados, estos se integraron en una frase canónica que resume la sintomatología del paciente en inglés clínico. Finalmente, dicha frase se incorporó dentro de una anamnesis sintética que incluye información demográfica esencial, como edad y

sexo, dando lugar al campo `note_text`. Este texto enriquecido constituye la representación final que posteriormente se transforma en embeddings y sirve como entrada para el modelo de clasificación diagnóstica.

El preprocesamiento de EVIDENCES consistió, por tanto, en un proceso estructurado de normalización semántica. Para cada paciente, se transformó la lista de evidencias identificando tanto los códigos binarios como aquellos que incluyen valores asociados. A partir del diccionario canónico, cada código fue traducido a su descripción clínica correspondiente. Por ejemplo, el código `E_91` se interpretó como `fever`, mientras que evidencias multivalor, como `E_54_@_V_161`, se tradujeron a expresiones completas como `moderate pain`. Al consolidar todas las descripciones generadas para un mismo paciente, se construyó una narrativa sintética breve y clínicamente coherente, que simula una anamnesis inicial y que incorpora, en su apertura, la edad y el sexo del paciente.

6.2.1. MODELOS DE EMBBINGS PARA ANAMNESIS SINTÉTICAS.

La vectorización del texto clínico se realizó utilizando modelos de lenguaje biomédico pre entrenados. En esta etapa, cada anamnesis sintética fue transformada en un vector numérico de 768 dimensiones mediante modelos de embeddings especializados en dominio clínico. Con el fin de identificar la representación semántica más adecuada para el modelamiento, se evaluaron cinco modelos biomédicos ampliamente utilizados para la generación de embeddings:

1. MedGemma: Modelo desarrollado por Google DeepMind y adaptado por Sentence-Transformers, entrenado específicamente en corpus clínicos y biomédicos.
2. Bio_ClinicalBERT: Modelo creado por investigadores del Harvard Medical School y MIT, entrenado sobre notas clínicas reales del dataset MIMIC-III y artículos biomédicos de PubMed.

3. SapBERT: Desarrollado por el Language Technology Lab de la Universidad de Cambridge, se basa en PubMedBERT y utiliza un enfoque de autoalineamiento semántico para mejorar la representación de conceptos médicos.
4. BioBERT: Modelo propuesto por el DMIS Lab de Korea University, entrenado sobre grandes colecciones de literatura científica biomédica incluyendo PubMed y PubMed Central.
5. SciBERT: Modelo desarrollado por el Allen Institute for AI, pre entrenado en más de un millón de artículos científicos.

Todos los vectores de embeddings resultantes (uno por paciente) se almacenaron en una matriz de características, de modo que pudieran ser reutilizados eficientemente en la siguiente etapa de modelado supervisado.

En cuanto a la variable objetivo, el diagnóstico confirmado del paciente (PATHOLOGY) ya venía codificado en el dataset con el nombre estandarizado de la enfermedad. Este campo se empleó con un label encoding como la etiqueta de clase para el aprendizaje supervisado, representando así un problema de clasificación multiclase (cada posible patología constituye una clase distinta a predecir). Adicionalmente, se respetó la partición de los datos proporcionada por DDxPlus, utilizando los subconjuntos predefinidos de entrenamiento y prueba incluidos en el lanzamiento oficial del dataset.

6.2.2. MODELOS DE CLASIFICACIÓN (MLP).

Para la modelación, se entrenó un perceptrón multicapa (MLP) denominado MLPDx, diseñado específicamente para clasificar el diagnóstico médico a partir del vector embebido de las anamnesis sintéticas antes mencionadas.

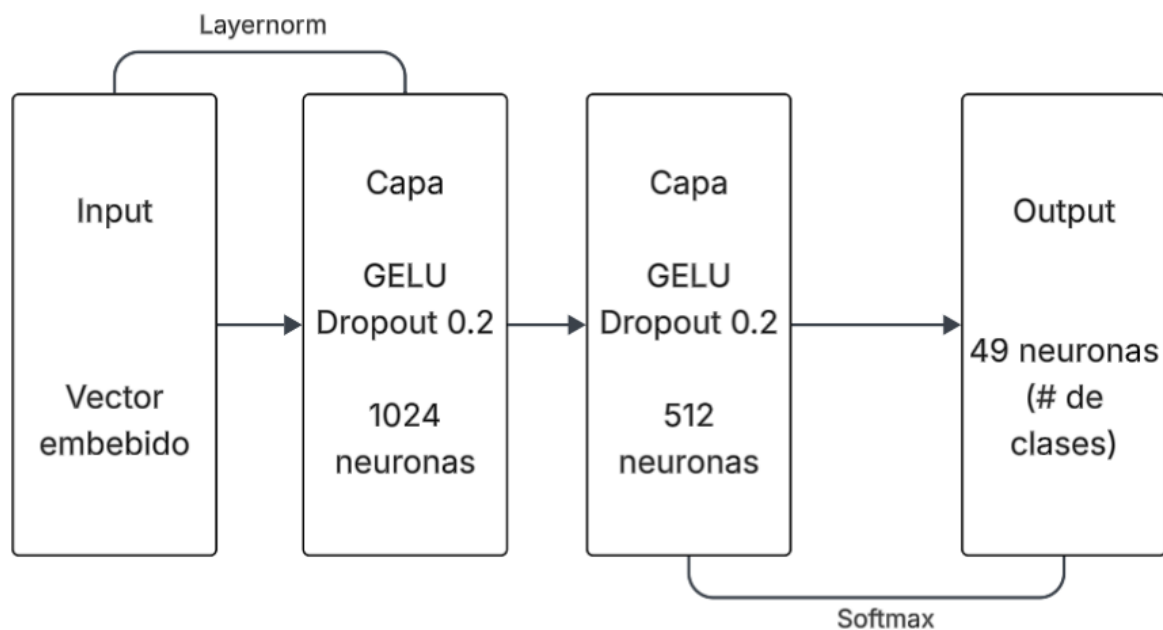


Figura 3. Arquitectura del perceptrón
 Nota. Fuente: Elaboración propia

La Figura 3 presenta de manera esquemática la arquitectura utilizada. El MLPDx está compuesto por una capa de entrada que recibe el vector embebido de 768 dimensiones, seguida de dos capas ocultas densas de 1024 y 512 neuronas, respectivamente, ambas con función de activación GELU, regularización mediante Dropout (0,2) y una normalización inicial LayerNorm para estabilizar el entrenamiento. Finalmente, la capa de salida cuenta con 49 neuronas, correspondientes al número total de diagnósticos distintos en el conjunto de datos, y produce los logits que posteriormente se transforman en probabilidades mediante una función softmax durante la inferencia.

La elección de esta arquitectura se sustenta en la literatura biomédica reciente, donde se han utilizado configuraciones de perceptrón multicapa con características muy similares para resolver problemas clínicos complejos. Park et al. (2025) emplean un MLP conformado por dos capas

ocultas de 1024 y 512 neuronas, complementado con Dropout, para predecir concentraciones de tacrolimus a partir de datos clínicos y de laboratorio. Sus resultados muestran que este tipo de arquitectura es capaz de modelar adecuadamente interacciones farmacocinéticas no lineales, superando incluso a métodos tradicionales como gradient boosting en escenarios de regresión clínica. Por su parte, Abduvaitov et al. (2025) integran capas densas finales de 1024 y 512 neuronas dentro de un pipeline basado en redes convolucionales para tareas de segmentación y clasificación de imágenes médicas en CT, MRI y radiografías de tórax. Estos autores evidencian que la inclusión de bloques densos profundos, acompañados de Dropout, potencia la capacidad del modelo para aprovechar representaciones de alto nivel y, al mismo tiempo, mitiga el sobreajuste en datos biomédicos de elevada complejidad. En conjunto, estos trabajos respaldan el uso de una arquitectura MLP con capas densas amplias, activaciones no lineales suaves y estrategias explícitas de regularización, constituyendo un equilibrio adecuado entre capacidad expresiva, estabilidad durante el entrenamiento y eficiencia computacional para la tarea de clasificación diagnóstica a partir de embeddings clínicos.

Finalmente, en concordancia con el tercer objetivo específico de esta investigación, se evaluó el desempeño del modelo utilizando métricas centradas en la relevancia clínica de sus clasificaciones. En lugar de limitarse a una exactitud global, se optó por una estrategia basada en el análisis de la cobertura de las predicciones principales, implementando las métricas top-1 y top-3 accuracy. Este enfoque es consistente con la literatura reciente en diagnóstico asistido por IA, donde se analiza si el diagnóstico verdadero se encuentra dentro de las primeras hipótesis propuestas por el sistema más que en una única predicción puntual. En este trabajo, la métrica top-1 se define como el porcentaje de casos en los que el diagnóstico verdadero del paciente coincide con la clase de mayor probabilidad predicha por el modelo, mientras que la top-3 mide

la proporción de ocasiones en que dicho diagnóstico se encuentra dentro de las tres primeras predicciones con mayor probabilidad. Este esquema de evaluación resulta especialmente pertinente en contextos clínicos donde, al igual que en los estudios de Tu et al. (2024) y Chishti et al. (2020), el sistema no se concibe como un emisor de un único dictamen categórico, sino como un generador de un conjunto acotado de hipótesis diagnósticas plausibles. Desde esta perspectiva, una top-3 accuracy elevada indica que el modelo es capaz de “pensar” en la enfermedad correcta dentro de sus primeras opciones, reproduciendo de forma aproximada el proceso de construcción de diagnósticos diferenciales que realizan los profesionales de la salud y aumentando su utilidad práctica como herramienta de apoyo a la toma de decisiones.

Como complemento a esta validación cuantitativa, se diseñó una prueba de estrés en un entorno simulado, con el propósito de evaluar la robustez del sistema frente a entradas reales. Esta prueba consistió en exponer el modelo a un conjunto de casos clínicos expresados en lenguaje natural por un profesional de la salud, sin intervención directa del desarrollador. Durante esta interacción controlada, el sistema fue desplegado dentro de un agente conversacional funcional, que recibió las descripciones de los síntomas, las procesó automáticamente y generó una primera impresión diagnóstica. Esta fase de evaluación buscó determinar la utilidad del sistema como herramienta de apoyo a la toma de decisiones durante la primera impresión diagnóstica, especialmente en escenarios de presión asistencial o de acceso limitado a especialistas, asegurando que su funcionamiento fuera coherente, interpretable y seguro en condiciones similares a las del entorno clínico real.

7. RESULTADOS.

7.1. RESULTADOS DEL OBJETIVO ESPECÍFICO 1: PREPARACIÓN Y ANÁLISIS EXPLORATORIO DEL CONJUNTO DE DATOS.

7.1.1. ANÁLISIS EXPLORATORIO DE DATOS (EDA).

El análisis exploratorio de datos (EDA, por sus siglas en inglés) constituye una etapa primordial en todo proceso de análisis, ya que permite comprender la estructura del conjunto de datos, evaluar la distribución de las variables y detectar posibles inconsistencias, patrones o sesgos que puedan influir en los resultados posteriores.

El presente análisis exploratorio busca caracterizar los datos desde una perspectiva descriptiva, evaluando la distribución de variables clave y sentando las bases para la posterior construcción de modelos predictivos.

7.1.2. CARACTERÍSTICAS DEMOGRÁFICAS DE LA MUESTRA.

La Figura 4 presenta la distribución de edades de la población incluida en el conjunto de datos, la variable edad, oscila entre 0 y 109 años, con una media cercana a los 39 años, como se refleja en la siguiente figura:

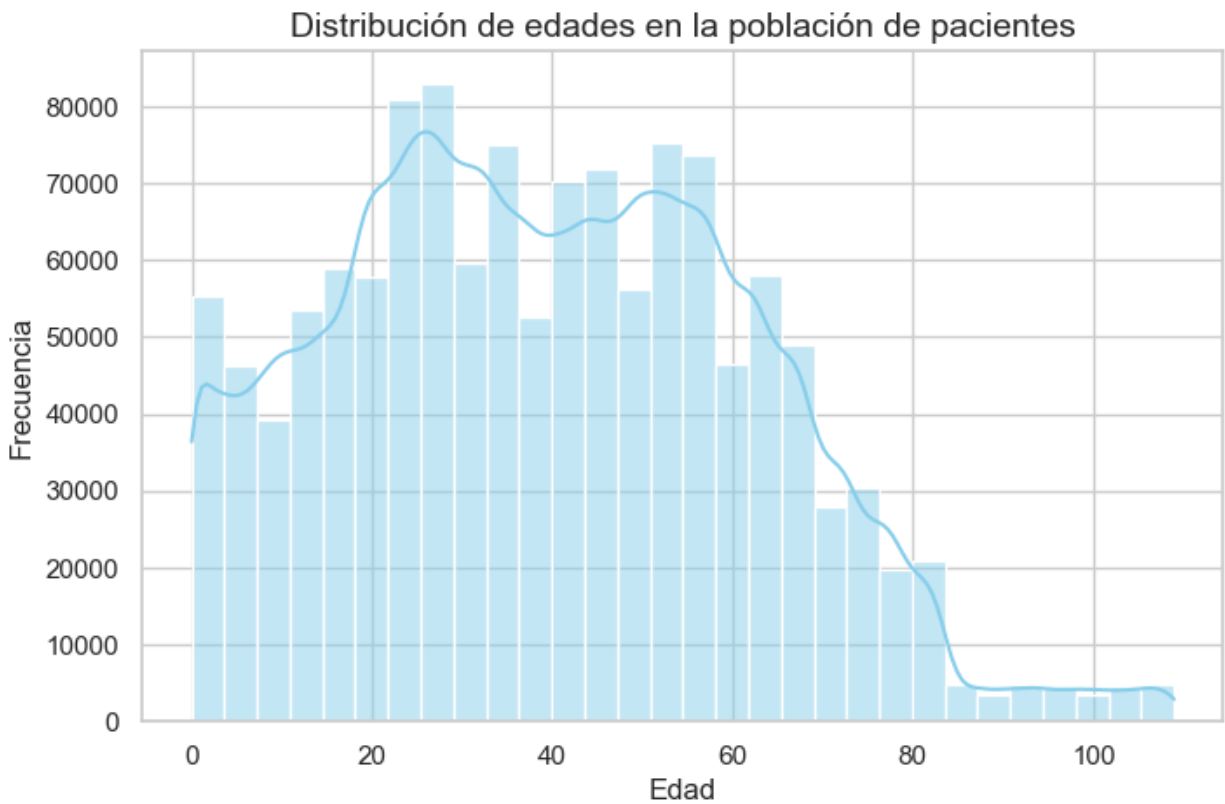


Figura 4. Distribución de edades en la población de pacientes
 Nota. Fuente: Elaboración propia

La Figura 5 muestra la distribución por la variable “SEX” a la cual hace referencia al sexo biológico de la población incluida en el dataset, evidenciando un equilibrio adecuado entre pacientes masculinos y femeninos. Aunque existe una ligera mayor representación de mujeres, la diferencia es mínima y no compromete la simetría del conjunto.

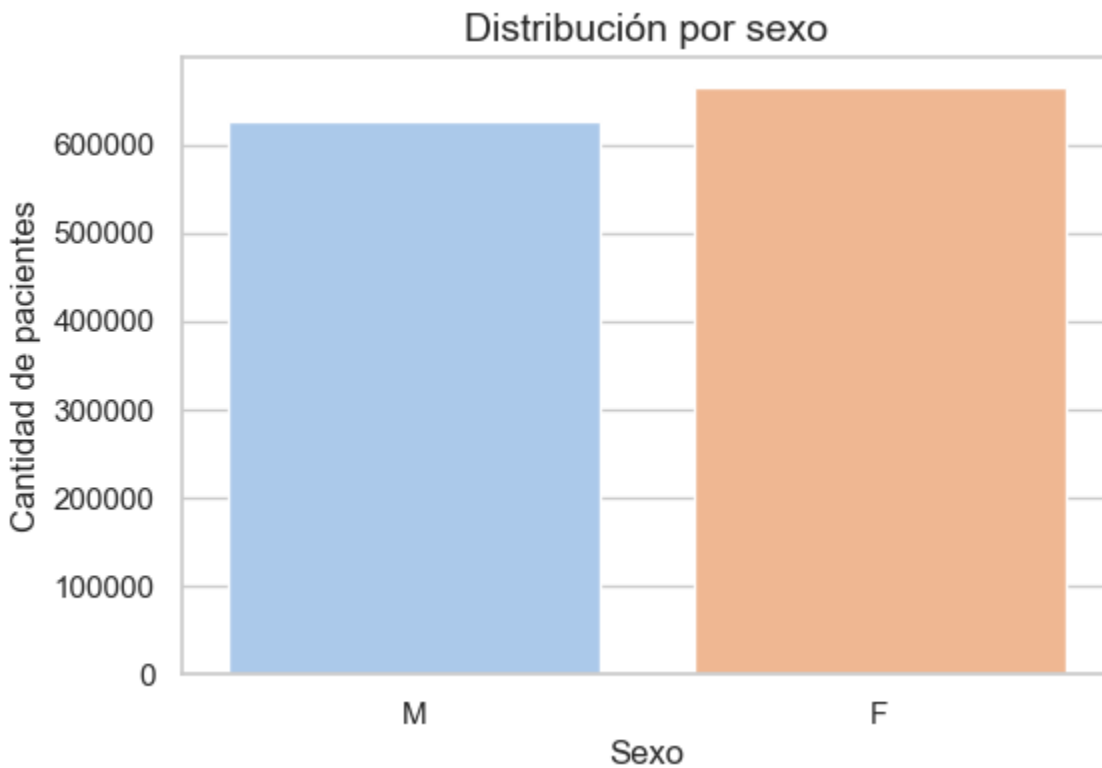


Figura 5. Distribución por sexo
Nota. Fuente: Elaboración propia

La Figura 6 compara la distribución de edades entre hombres y mujeres mediante un diagrama de cajas. Ambos grupos presentan patrones muy similares, con medianas cercanas a los 35–40 años y rangos intercuartílicos prácticamente equivalentes. También se observa una amplia variabilidad en ambas distribuciones, con edades que oscilan desde la infancia hasta más de 100 años, lo que confirma la diversidad etaria del conjunto de datos. La presencia de pocos valores atípicos aislados en mujeres refleja casos excepcionales de edades extremas, pero no altera de forma significativa la simetría general.

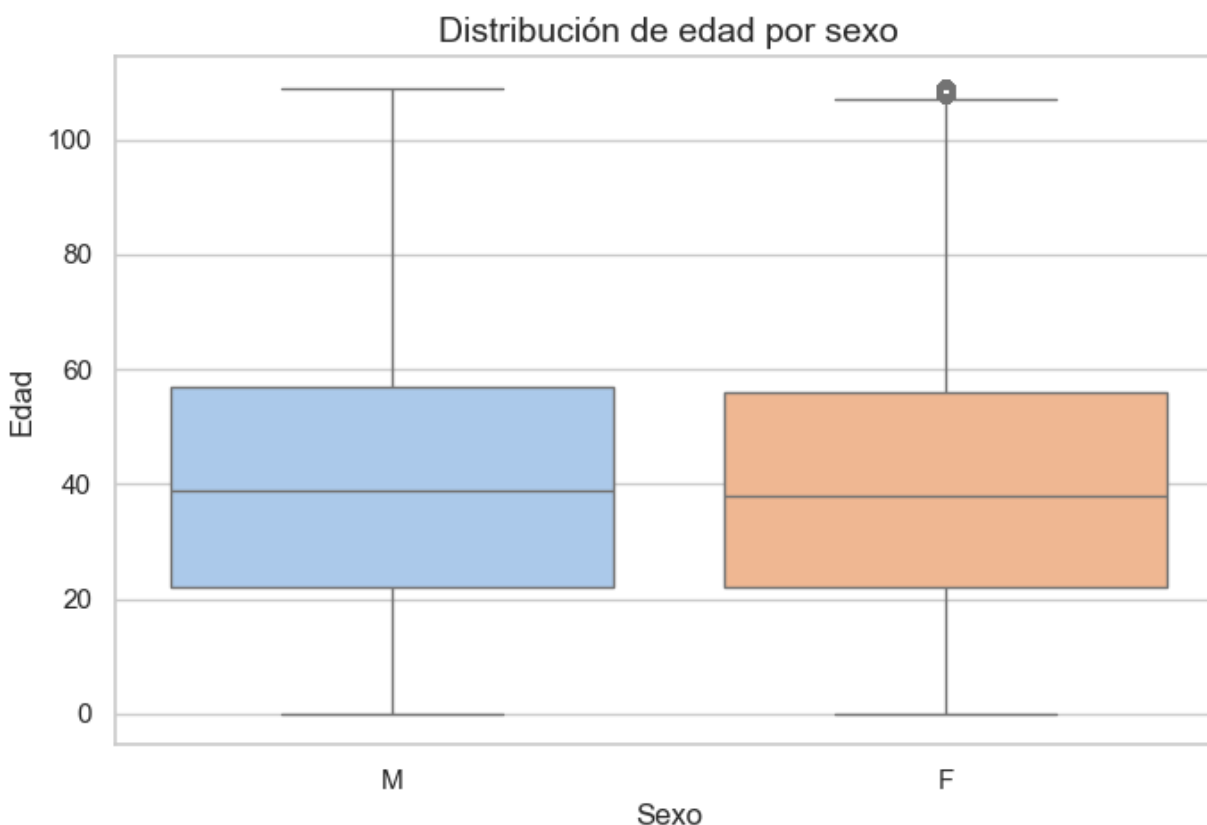


Figura 6. Distribución de edad por sexo
 Nota. Fuente: Elaboración propia

7.1.3. DISTRIBUCIÓN DE DIAGNÓSTICOS Y EVIDENCIAS CLÍNICAS.

La Figura 7 presenta los diez diagnósticos más frecuentes en la base de datos, evidenciando una predominancia de patologías respiratorias e infecciosas. Las infecciones de vías respiratorias superiores (URTI) y la faringitis viral encabezan la lista con la mayor cantidad de casos, seguidas por condiciones como la anemia y la infección inicial por VIH. El resto de diagnósticos muestra frecuencias más uniformes, incluyendo anafilaxia, edema localizado, embolismo pulmonar, influenza, bronquitis y enfermedad por reflujo gastroesofágico (GERD).

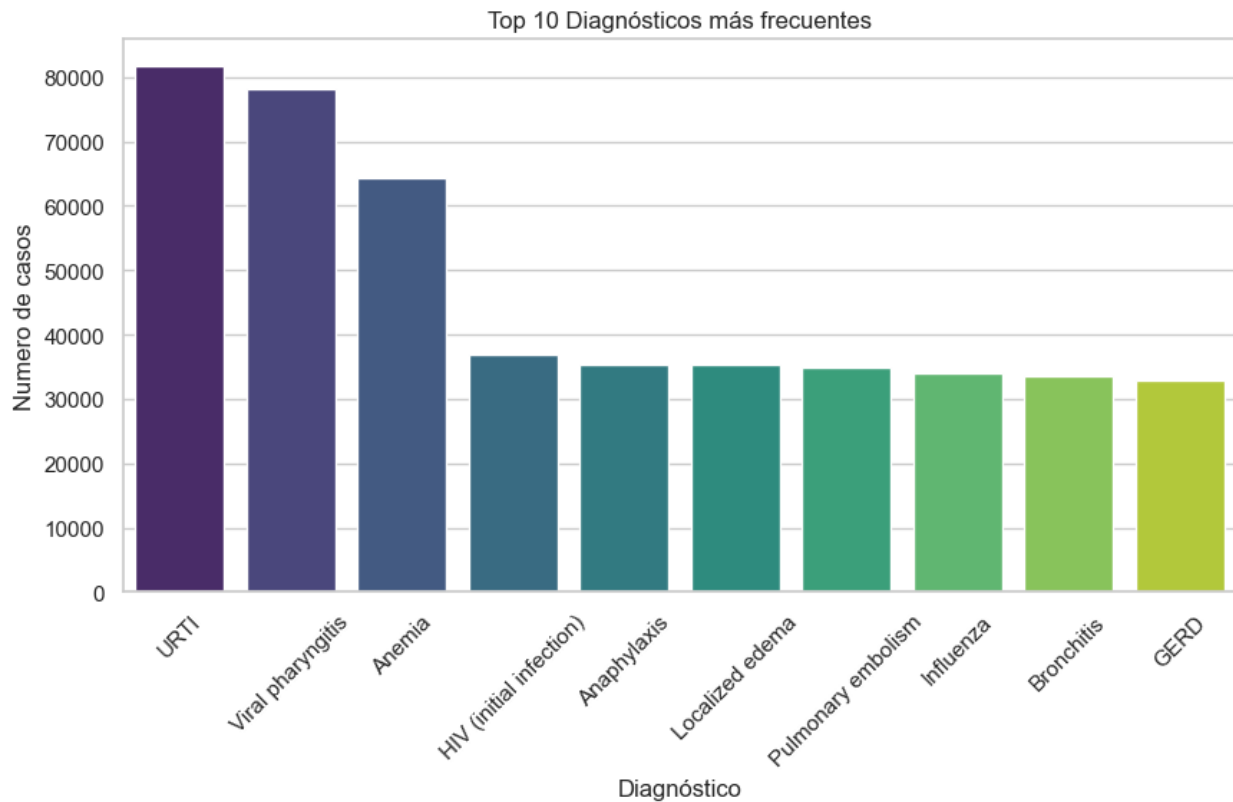


Figura 7. Top 10 diagnósticos más frecuentes
 Nota. Fuente: Elaboración propia

La Figura 8 muestra la distribución del sexo para los diez diagnósticos más frecuentes del conjunto de datos. En general, se observa un patrón equilibrado entre hombres y mujeres en la mayoría de las patologías, aunque algunas presentan ligeras variaciones. En diagnósticos como URTI, faringitis viral y anemia, la frecuencia es marginalmente mayor en mujeres, mientras que en el resto de condiciones las diferencias por sexo son mínimas. En conjunto, esta distribución sugiere que la prevalencia de los principales diagnósticos es relativamente homogénea entre ambos grupos, lo cual reduce la probabilidad de sesgos asociados al sexo en la etapa de modelamiento.

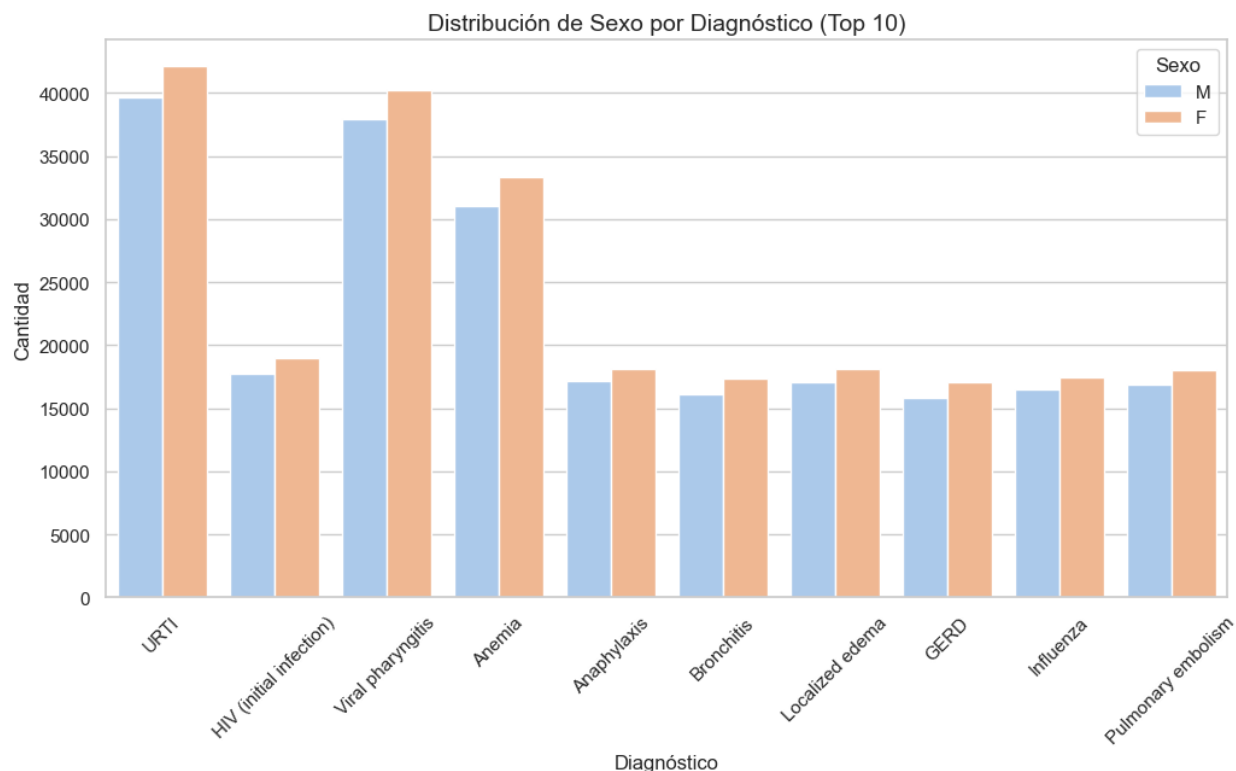


Figura 8. Distribucion de sexo por diagnostico (Top 10)
 Nota. Fuente: Elaboración propia

La Figura 9 muestra la distribución del número de síntomas reportados por paciente en la variable EVIDENCES. La mayoría de los casos presenta entre 10 y 25 síntomas, con un pico marcado alrededor de las 18 a 20 evidencias, lo que indica que los registros suelen contener suficiente información clínica para caracterizar el cuadro del paciente. La distribución es amplia y ligeramente sesgada hacia la derecha, con presencia de casos que superan los 30 o 40 síntomas, los cuales representan escenarios clínicos más complejos o altamente inespecíficos. En conjunto, esta variabilidad en el tamaño de las listas sintomáticas refleja la heterogeneidad del dataset y proporciona una base rica para el modelamiento predictivo, al capturar tanto situaciones simples como casos con múltiples manifestaciones clínicas.

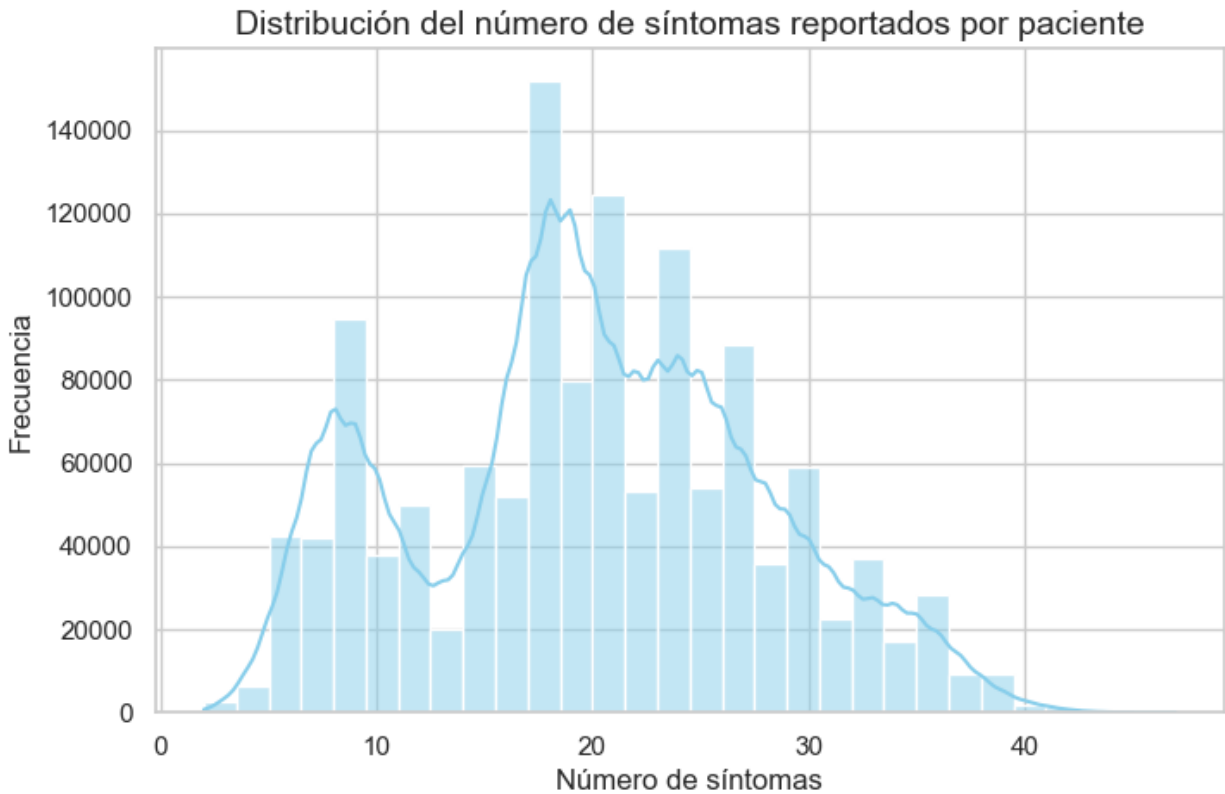


Figura 9. Distribución del número de síntomas reportados por paciente
 Nota. Fuente: Elaboración propia

7.2. RESULTADOS DEL OBJETIVO ESPECÍFICO 2: DESEMPEÑO DEL MODELO DE CLASIFICACIÓN DIAGNÓSTICA.

En concordancia con el segundo objetivo específico y con las fases de preparación de los datos y modelado de la metodología CRISP-DM, se evaluó el desempeño del clasificador diagnóstico entrenado a partir de las anamnesis sintéticas (note_text) generadas en la etapa previa. Cada nota clínica fue transformada en un vector denso mediante cinco modelos de embeddings biomédicos pre entrenados, y estos vectores sirvieron como entradas para un perceptrón multicapa (MLPDx) encargado de predecir el diagnóstico confirmado del paciente. El problema se formuló como una tarea de clasificación multiclase, donde cada patología corresponde a una clase distinta.

La métrica principal utilizada fue la precisión Top-1, definida como la proporción de casos en los que el diagnóstico verdadero coincide con la clase de mayor probabilidad predicha por el modelo. Esta métrica resulta especialmente exigente, ya que no considera diagnósticos alternativos plausibles, y permite valorar la capacidad del sistema para acertar en su primera sugerencia diagnóstica.

La Figura 10 resume los resultados obtenidos para cada modelo de embeddings clínicos evaluado. Tres de ellos (Embedding-Gemma-300M-Medical, Bio_ClinicalBERT y SapBERT) alcanzaron una precisión Top-1 aproximada a 99,2 %, mientras que BioBERT obtuvo un 97,2 % y SciBERT un 95,5 %

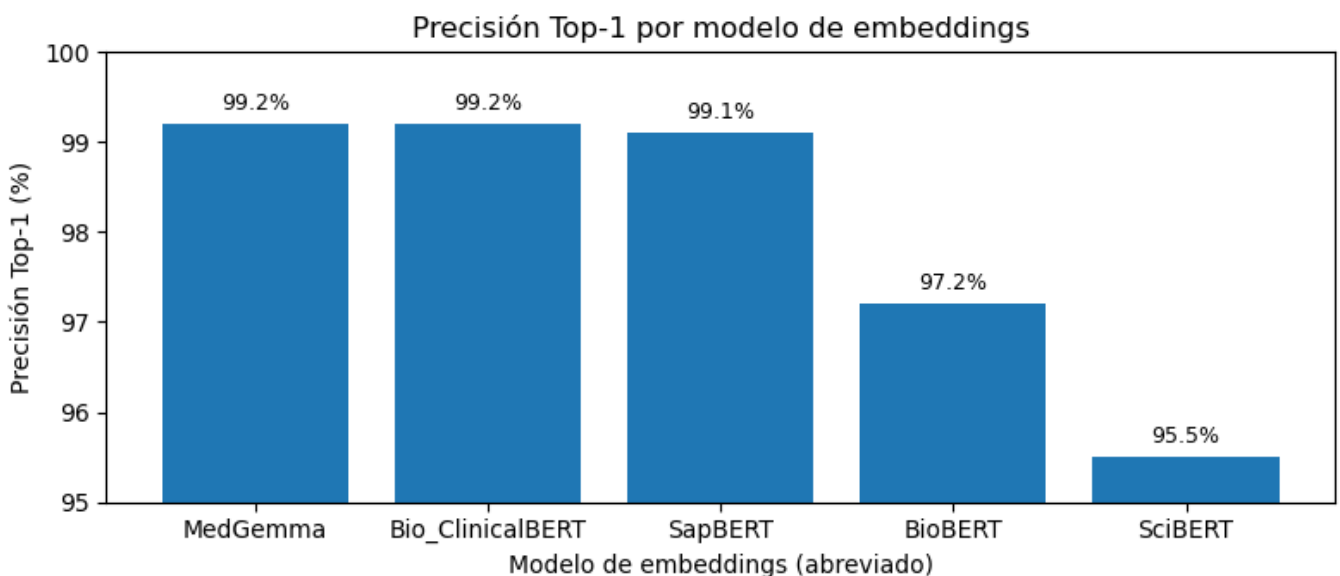


Figura 10. Precisión Top-1 por modelo de embeddings
Nota. Fuente: Elaboración propia

Estos hallazgos muestran que los modelos entrenados o afinados específicamente sobre corpus biomédicos y clínicos presentan un rendimiento claramente superior frente a modelos científicos más generales, lo que confirma la importancia del alineamiento de dominio para capturar las relaciones semánticas entre síntomas y diagnósticos. En relación con la precisión Top-3, todos los modelos evaluados alcanzaron un desempeño prácticamente idéntico, con valores cercanos al

99,9 %. Dado que no se observan diferencias relevantes entre los resultados para esta métrica, su representación gráfica no aporta información adicional para la comparación; por este motivo, dichos resultados se presentan únicamente de forma textual y el análisis detallado se centra en la precisión Top-1.

Al poner en perspectiva estos resultados frente a la literatura especializada en modelos diagnósticos que emplean bases de datos distintas a DDxPlus se evidencia que Tu et al. (2024) evaluaron un sistema de diálogo clínico basado en modelos de lenguaje (AMIE) mediante métricas top-k, mostrando que el modelo rivaliza a médicos de atención primaria en la exactitud de las listas diferenciales cuando se consideran las primeras posiciones del ranking diagnóstico. De manera similar, Chishti et al. (2020) validaron un modelo probabilístico de diagnóstico diferencial frente a médicos humanos utilizando la top-3 accuracy como métrica principal y evidenciaron que el sistema alcanzaba una cobertura del 85 % del diagnóstico correcto dentro de sus tres primeras sugerencias, frente a aproximadamente un 67 % en el grupo de especialistas. Estos estudios respaldan el uso de métricas top-k como indicadores adecuados del valor clínico de los modelos diagnósticos, especialmente cuando se espera que la IA funcione como apoyo al razonamiento diferencial más que como reemplazo del juicio médico.

Sin embargo, al comparar estos resultados con otros trabajos que utilizan el mismo conjunto de datos sintéticos DDxPlus, es posible contextualizar mejor el desempeño obtenido. La Tabla 2 resume algunos estudios recientes, indicando los modelos empleados y el rendimiento reportado. Es importante señalar que, en todos los casos, las métricas de la tabla corresponden al desempeño del modelo baseline de Action-Aware Reinforcement Learning Classifier (AARLC) evaluado por cada autor sobre DDxPlus..

Autores	Año	Artículo	Modelo	Desempeño
Fansi Tchango et al.	2022	DDXPlus: A new Dataset for Medical Automatic Diagnosis	Action-Aware Reinforcement Learning Classifier	98.9%
Yang et al.	2024	Adaptive Top-K Algorithm for Medical Conversational Diagnostic Model	Adaptive Top-K Algorithm	99.9%
Alam et al.	2023	DDxT: Deep Generative Transformer Models for Differential Diagnosis	Transformer	99.2%

Tabla 2. Desempeño de modelos AARLC en DDXPlus.

Como se observa, los valores de desempeño reportados en los artículos mencionados (Basados principalmente en modelos de aprendizaje por refuerzo para la selección y clasificación de diagnósticos) se sitúan aproximadamente entre el 98,9 % y el 99,9 %. En contraste, los modelos de embeddings evaluados en el desarrollo de esta tesis, los cuales se apoyan en una arquitectura de red neuronal con aprendizaje supervisado, alcanzan valores de desempeño cercanos al 99,2 % de precisión Top-1. Este resultado es consistente con el intervalo reportado en la literatura e, incluso, ligeramente superior a algunos de los valores informados en determinados estudios.

Esta comparación sugiere que un enfoque basado en redes neuronales entrenadas con representaciones semánticas de alta calidad puede alcanzar niveles de desempeño equiparables a los logrados por agentes más complejos orientados por aprendizaje por refuerzo sobre el mismo conjunto de datos.

7.3. RESULTADOS DEL OBJETIVO ESPECÍFICO 3: PRUEBA DE ESTRÉS Y EVALUACIÓN DEL AGENTE CONVERSACIONAL.

Como parte del tercer objetivo específico y de las fases de evaluación y despliegue de la metodología CRISP-DM, se diseñó una interfaz conversacional que integra el modelo diagnóstico previamente seleccionado. Esta interfaz permite que el usuario ingrese una descripción sintomática en lenguaje natural y reciba, en respuesta, una sugerencia diagnóstica generada automáticamente por el sistema. El diseño se enfocó en lograr una interacción clara, sencilla y compatible con los tiempos de uso propios de una consulta clínica.

A continuación, se presentan las capturas de la interfaz prototipo propuesta para el agente conversacional (Figuras 11–14), donde se ilustra el flujo básico de interacción entre el usuario y el sistema.



Figura 11. Interfaz inicial
Nota. Fuente: Elaboración propia

En la Figura 11 se presenta el módulo principal de interacción con el agente, diseñado para permitir la introducción libre de la descripción de síntomas o anamnesis reportada por el paciente. La interfaz dispone de un campo de texto amplio en el que el usuario puede detallar los síntomas en lenguaje natural, seguido de un botón denominado “Ejecutar inferencia en todos los

modelos activos”, el cual actúa como disparador para procesar la entrada mediante los modelos de embeddings previamente cargados. En el panel lateral izquierdo se visualiza el conjunto de modelos disponibles, los cuales pueden activarse o desactivarse de forma individual, permitiendo comparar su comportamiento bajo una misma descripción clínica. Finalmente, la interfaz incorpora un control deslizante para ajustar dinámicamente el valor de Top-k, definiendo cuántos diagnósticos sugeridos serán presentados en los resultados, lo que facilita adaptar el nivel de detalle según la complejidad de la anamnesis ingresada.

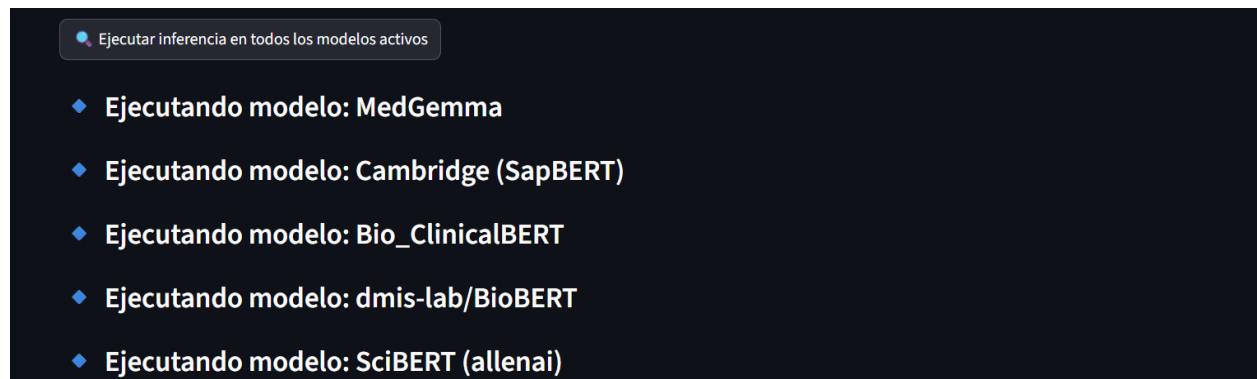


Figura 12. Cargue modelos
Nota. Fuente: Elaboración propia

En la Figura 12 se muestran los mensajes de ejecución que la interfaz despliega una vez iniciado el proceso de inferencia. Estos avisos indican, de manera secuencial, el cargue y la activación de cada uno de los modelos de embeddings seleccionados por el usuario, lo que permite visualizar el avance interno del sistema y proporciona retroalimentación inmediata sobre el estado del proceso

2 Resultados comparativos por modelo ↗	
♦ MedGemma	
Diagnóstico	Probabilidad
Croup	1.000
Bronchitis	0.000
♦ Cambridge (SapBERT)	
Diagnóstico	Probabilidad
Epiglottitis	0.970
Unstable angina	0.029
♦ Bio_ClinicalBERT	
Diagnóstico	Probabilidad
Unstable angina	0.998
Spontaneous rib fracture	0.002

Figura 13. Resultados 1.
 Nota. Fuente: Elaboración propia

♦ dmis-lab/BioBERT	
Diagnóstico	Probabilidad
Croup	0.998
Anaphylaxis	0.001
♦ SciBERT (allenai)	
Diagnóstico	Probabilidad
Scombroid food poisoning	0.803
Croup	0.172

Figura 14. Resultados 2.
 Nota. Fuente: Elaboración propia

En las Figuras 13 y 14 se presentan los resultados generados por el agente diagnóstico tras procesar la anamnesis ingresada y aplicar las configuraciones seleccionadas por el usuario. La interfaz despliega, para cada modelo de embeddings activo, los diagnósticos con mayor probabilidad según el valor de Top-k definido, junto con sus respectivas estimaciones numéricas. Esta visualización comparativa permite observar de forma clara cómo cada modelo prioriza distintas hipótesis diagnósticas a partir del mismo conjunto de síntomas, evidenciando diferencias en sensibilidad y coherencia clínica.

Una vez definida y probada la interfaz, se procedió a realizar una prueba de estrés orientada a evaluar el desempeño funcional del sistema en un escenario de uso libre. Para ello, un profesional de la salud interactuó con el agente utilizando descripciones sintomáticas redactadas de forma espontánea, sin seguir la estructura estandarizada del dataset original. Este ejercicio permitió aproximarse a las condiciones reales de un entorno clínico, donde los pacientes expresan sus síntomas con niveles variables de detalle, ambigüedad y orden narrativo.

El objetivo de esta prueba de estrés era el de evaluar aspectos fundamentales del comportamiento del agente como:

1. Coherencia clínica en las primeras impresiones diagnósticas sugeridas por el agente.
2. Consistencia en el modelo ante variaciones en la formulación de los síntomas

Los resultados preliminares mostraron que el agente conversacional fue capaz de producir diagnósticos plausibles y alineados con la sintomatología general del paciente y de generar una primera impresión diagnóstica coherente con la sintomatología general simulada por el profesional de la salud, identificando adecuadamente categorías clínicas amplias como afecciones respiratorias, infecciosas, cardíacas o alérgicas. No obstante, debido a que el dataset utilizado para el entrenamiento incluye únicamente 49 diagnósticos principales y está construido a partir de un perfil poblacional específico (Personas residentes de nueva york), el sistema presentó dificultades para aproximarse con precisión al diagnóstico exacto esperado en algunos casos. Esta limitación refleja tanto el alcance restringido del conjunto de clases disponibles como la falta de representatividad clínica de ciertos escenarios locales. Aun así, el comportamiento observado indica que el prototipo del agente de IA es funcional y técnicamente viable, y constituye una base sólida para fases de validación posteriores con textos clínicos reales o entornos supervisados más cercanos a la práctica asistencial.

Más allá de la precisión cuantitativa, esta fase permitió valorar la viabilidad práctica del agente de IA como herramienta de apoyo a la primera impresión diagnóstica. La interacción conversacional evidenció que el modelo puede integrarse como un asistente cognitivo que organiza la información sintomática, sugiere primeras impresiones diagnósticas y facilita la exploración de alternativas, siempre bajo la supervisión del profesional de la salud. Estos resultados cualitativos respaldan el cumplimiento del tercer objetivo específico y muestran que el sistema, aunque aún requiere validaciones adicionales sobre textos clínicos reales, constituye un prototipo funcional con potencial para ser refinado y evaluado en estudios futuros de mayor escala.

8. CONCLUSIONES.

El trabajo desarrollado demuestra la viabilidad técnica de un agente de inteligencia artificial para apoyar la formulación de una primera impresión diagnóstica a partir de descripciones sintomáticas en lenguaje natural. La metodología CRISP-DM permitió articular de forma estructurada la comprensión del problema clínico, el análisis exploratorio, la preparación de datos, el modelado, la evaluación y un despliegue preliminar del sistema, consolidando un flujo de trabajo reproducible y adaptable a otros contextos asistenciales.

En relación con los objetivos específicos, el análisis del conjunto de datos sintéticos DDxPlus permitió caracterizar la estructura y distribución de variables demográficas, clínicas y sintomáticas, confirmando su idoneidad como insumo para modelos de diagnóstico asistido. El proceso de normalización semántica transformó listas de evidencias en anamnesis sintéticas legibles (note_text), integrando edad, sexo y síntomas en un texto clínico estructurado que facilitó el uso de modelos de embeddings entrenados con lenguaje biomédico. A partir de estas representaciones, la combinación de embeddings clínicos y el perceptrón multicapa MLPDx

alcanzó desempeños muy competitivos (precisión Top-1 cercana al 99,2 % y Top-3 del 99,9 %), evidenciando la capacidad de los modelos especializados para capturar la semántica de las narrativas clínicas.

Finalmente, la integración del modelo en un agente capaz de interactuar con el profesional de la salud, razonar sobre la información disponible y sugerir primeras impresiones diagnósticas permitió, a través de una prueba de estrés con descripciones sintomáticas redactadas libremente, comprobar que el sistema genera hipótesis plausibles en grandes categorías clínicas. No obstante, su rendimiento se ve condicionado por el uso exclusivo de datos sintéticos y la cobertura limitada a 49 diagnósticos. En conjunto, se concluye que el agente cumple su propósito como prototipo de apoyo cognitivo al profesional, pero no como sistema autónomo de decisión clínica. Estos resultados ponen de relieve la necesidad de futuras líneas de trabajo orientadas a ampliar la cobertura diagnóstica, incorporar datos reales, reforzar los mecanismos de explicabilidad y controlar posibles sesgos, recordando que un alto desempeño cuantitativo en entornos controlados no garantiza, por sí solo, una efectividad equivalente en la práctica clínica cotidiana.

9. REFERENCIAS

- Abduvaitov, A., Buriboev, A. S., Sultanov, D., Buriboev, S., Yusupov, O., Kilichov, J., & Choi, A. J. (2025). Enhancing medical image segmentation and classification using a fuzzy-driven method. *Sensors*, 25(18), 5931.
- Agency for Healthcare Research and Quality. (2020). Making healthcare safer III: A critical analysis of existing and emerging patient safety practices. Chapter 1: Diagnostic errors. Agency for Healthcare Research and Quality (US). <https://www.ncbi.nlm.nih.gov/books/NBK555525/>
- Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med*. 2021 Apr 7;4(1):65. doi: 10.1038/s41746-021-00438-z. PMID: 33828217; PMCID: PMC8027892.
- Alam, M. M., Raff, E., Oates, T., & Matuszek, C. (2023). DDxT: Deep generative transformer models for differential diagnosis. arXiv. <https://doi.org/10.48550/arXiv.2312.01242>
- Báez, C., Poblete, J., & López, J. (2022). Procesamiento de lenguaje natural para el análisis de textos clínicos en español: Estudio de caso en listas de espera. *Revista Chilena de Salud Pública*, 26(1), 34–43.
- Bartlett, J. D., Keith, M. S., Sudharshan, L., & Snedecor, S. J. (2015). Associations between signs and symptoms of dry eye disease: A systematic review. *Clinical Ophthalmology*, 9, 1719–1730. <https://doi.org/10.2147/OPTH.S89700>
- Chishti, S., Jaggi, K. R., Saini, A., Agarwal, G., & Ranjan, A. (2020). Artificial intelligence–based differential diagnosis: Development and validation of a probabilistic model to address lack of large-scale clinical datasets. *Journal of Medical Internet Research*, 22(4), e17550.
- Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, 11, 70.

- Diez, E., & de la Coteria, M. (2010). El diagnóstico médico, problema crucial en el siglo XXI. *Medisur*, 8(2).
- D'Souza, R. F., Amanullah, S., Mathew, M., & Surapaneni, K. M. (2023). Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry*, 89, 103770. <https://doi.org/10.1016/j.ajp.2023.103770>
- El Arab, R. A., Abu-Mahfouz, M. S., Abuadas, F. H., Alzghoul, H., Almari, M., Ghannam, A., & Seweid, M. M. (2025). Bridging the Gap: From AI Success in Clinical Trials to Real-World Healthcare Implementation—A Narrative Review. *Healthcare*, 13(7), 701. <https://doi.org/10.3390/healthcare13070701>
- Fansi Tchango, A., Wen, Z., Goel, R., Martel, J., & Ghosn, J. (2022). DDXPlus: A new dataset for automatic medical diagnosis. *arXiv*. <https://doi.org/10.48550/arXiv.2205.09148>
- Graber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf*. 2013 Oct;22 Suppl 2(Suppl 2):ii21-ii27. doi: 10.1136/bmjqs-2012-001615. Epub 2013 Jun 15. PMID: 23771902; PMCID: PMC3786666.
- Jeong J, Kim S, Pan L, Hwang D, Kim D, Choi J, Kwon Y, Yi P, Jeong J, Yoo SJ. Reducing the workload of medical diagnosis through artificial intelligence: A narrative review. *Medicine (Baltimore)*. 2025 Feb 7;104(6):e41470.
- Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. *Cureus*. 2023 Aug 10;15(8):e43262. doi: 10.7759/cureus.43262. PMID: 37692617; PMCID: PMC10492220.
- Kaur, Simarjeet & Singla, Jimmy & Nkenyereye, Lewis & Jha, Sudan & Prashar, Deepak & Joshi, Gyanendra Prasad & El-Sappagh, Shaker & Islam, Md & Islam, S. M. Riazul. (2020). Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. *IEEE Access*. PP. 1-1. 10.1109/ACCESS.2020.3042273.
- Newman-Toker, D. E., Nassery, N., Schaffer, A. C., Yu-Moe, C. W., Clemens, K. D., Natesan, S., Fanai, M., Moy, E., & Segal, J. B. (2024). Burden of serious harms from diagnostic error in the USA. *BMJ Quality & Safety*, 33(2), 109–120. <https://doi.org/10.1136/bmjqs-2021-014130>.

- Park, D. J., Kim, M., Cho, H., Kim, J. S., Hwang, J., & Lee, J. (2025). Machine learning-based algorithm for tacrolimus dose optimization in hospitalized kidney transplant patients. *Diagnostics*, 15(23), 2948.
- Rosendal, M., Jarbøl, D. E., Pedersen, A. F., & Andersen, R. S. (2013). Multiple perspectives on symptom interpretation in primary care research. *BMC Family Practice*, 14, 167. <https://doi.org/10.1186/1471-2296-14-167>
- Shamout F, Zhu T, Clifton DA. Machine Learning for Clinical Outcome Prediction. *IEEE Rev Biomed Eng.* 2021;14:116-126. doi: 10.1109/RBME.2020.3007816. Epub 2021 Jan 22. PMID: 32746368.
- Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, 19, 64. <https://doi.org/10.1186/s12874-019-0681-4>
- Singh, H., & Weingart, S. N. (2009). Diagnostic errors in ambulatory care: Dimensions and preventive strategies. *Advances in Health Sciences Education*, 14(Suppl 1), 57–61. <https://doi.org/10.1007/s10459-009-9177-z>
- Tu, T., McDuff, D. J., Liu, S., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 637, 115–123. <https://doi.org/10.1038/s41586-025-08866-7>
- Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. *Visc Med.* 2020 Dec;36(6):443-449. doi: 10.1159/000511930. Epub 2020 Oct 12. PMID: 33442551; PMCID: PMC7768160.
- Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y.-H., Shavick, A., ... Dobson, R. J. B. (2022). A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *npj Digital Medicine*, 5, 186.
- Yang, Y., Zhang, G., Wu, Y., Zhao, Z., & Fu, Y. (2024). Adaptive Top-K algorithm for medical conversational diagnostic model. *Entropy*, 26(9), 740. <https://doi.org/10.3390/e26090740>