UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MODELO TRANSFORMER PARA DIAGNOSTICAR CARDIOPATÍA ISQUÉMICA CRÓNICA DERIVADA DE DIABETES MELLITUS TIPO 2

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ORLANDO FLORES CUSTODIO

BAJO LA DIRECCIÓN DE:

DR. PABLO PANCARDO GARCÍA

CUNDUACÁN, TABASCO, A: SEPTIEMBRE 2025

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MODELO TRANSFORMER PARA DIAGNOSTICAR CARDIOPATÍA ISQUÉMICA CRÓNICA DERIVADA DE DIABETES MELLITUS TIPO 2

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ORLANDO FLORES CUSTODIO

BAJO LA DIRECCIÓN DE:

DR. PABLO PANCARDO GARCÍA

CUNDUACÁN, TABASCO, A: SEPTIEMBRE 2025

Declaración de Autoría y Originalidad

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco, a 01 de Septiembre de 2025.

Estudiante: Orlando Flores Custodio

UJAT
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

2025 AÑO DE LA Mujer Indígena

Cunduacán, Tabasco, a 09 de julio de 2025
Oficio No. 1202/2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Orlando Flores Custodio
Egresado de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Modelo transformer para diagnosticar cardiopatía isquémica crónica derivada de diabetes mellitus tipo 2"***, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

Atentamente

MTE. Óscar Alberto González González
Director

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

M.T.E. OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda. C.P. 86690
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Cunduacán, Tabasco, a 01 de Septiembre de 2025.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Modelo transformer para diagnosticar cardiopatía isquémica crónica derivada de Diabetes Mellitus tipo 2**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Orlando Flores Custodio

TESTIGOS

Dr. Pablo Pancardo García

Dr. José Adán Hernández Nolasco

Dedicatoria

- A Dios, por ser siempre generoso conmigo, porque hasta aquí nos ha ayudado Jehová.
- A mis padres, que siempre han tenido confianza en mí.
- A cada uno de los maestros que tuvieron a bien transmitirme parte de su conocimiento.
- A mi director de tesis, quien en cada momento tuvo las palabras acertadas para mi aprendizaje.
- A mí, por creer en que podría lograrlo, y luchar por ello.
- Especialmente a mi esposa Clarisa Fernanda Sánchez Blé, por siempre dar palabras de aliento y decirme que, aún en los peores momentos, no olvide sonreír.

Índice general

Índice de figuras

Índice de tablas

Resumen

La cardiopatía isquémica crónica (CIC) constituye una complicación frecuente en pacientes con diabetes mellitus tipo 2 (DM2), y su diagnóstico oportuno resulta fundamental para mejorar el tratamiento y pronóstico del paciente. Sin embargo, el diagnóstico se dificulta por la presencia atípica de síntomas y la diversidad de factores de riesgo. Esta investigación presenta el desarrollo, implementación, optimización y evaluación de un modelo de red neuronal tipo Transformer adaptado para datos tabulares (TabTransformer), capaz de diagnosticar CIC a partir de registros clínicos reales de pacientes con DM2.

Se utilizó un conjunto de datos con información longitudinal de pacientes mexicanos atendidos en el Instituto Mexicano del Seguro Social de Michoacán, México durante el periodo comprendido entre los años 2005 y 2020. El modelo alcanzó un rendimiento destacable con un *accuracy* de 87.72 %, un *recall* de 90.16 % y un valor de 0.9434 para el área bajo la curva ROC. Se priorizó especialmente el *recall* como métrica principal, aspecto crítico en el diagnóstico de enfermedades cardiovasculares.

Mediante un proceso riguroso de selección de características e implementando un algoritmo genético adaptativo para la optimización de hiperparámetros, se identificaron 18 variables determinantes para el diagnóstico. El modelo TabTransformer superó significativamente a los enfoques tradicionales de aprendizaje automático como regresión logística, k-vecinos más cercanos y árboles de decisión, demostrando una mayor capacidad para capturar relaciones complejas entre variables clínicas gracias a su mecanismo de atención. Los resultados confirman la efectividad del enfoque propuesto para el diagnóstico de CIC en pacientes diabéticos.

Palabras clave: Inteligencia Artificial, Redes neuronales, Aprendizaje automático.

Abstract

Chronic ischemic heart disease (CIHD) is a common complication in patients with type 2 diabetes mellitus (T2DM), and its timely diagnosis is essential to improve patient treatment and prognosis. However, diagnosis is hampered by the atypical presentation of symptoms and the diversity of risk factors. This research presents the development, implementation, optimization, and evaluation of a Transformer-type neural network model adapted for tabular data (TabTransformer), capable of diagnosing CIHD from real clinical records of patients with T2DM.

A dataset with longitudinal information from Mexican patients treated at the Mexican Social Security Institute of Michoacán, Mexico during the period between 2005 and 2020 was used. The model achieved remarkable performance with an accuracy of 87.72 %, a recall of 90.16 %, and a value of 0.9434 for the area under the ROC curve. Recall was especially prioritized as the main metric, a critical aspect in the diagnosis of cardiovascular diseases.

Through a rigorous feature selection process and implementing an adaptive genetic algorithm for hyperparameter optimization, 18 determinant variables for diagnosis were identified. The TabTransformer model significantly outperformed traditional machine learning approaches such as logistic regression, k-nearest neighbors, and decision trees, demonstrating a greater ability to capture complex relationships between clinical variables thanks to its attention mechanism. The results confirm the effectiveness of the proposed approach for the diagnosis of CIHD in diabetic patients.

Keywords: Artificial Intelligence, Neural Networks, Machine Learning.

Capítulo 1

Generalidades

1.1. Introducción

El diagnóstico de problemas de salud comenzó a ser formal con los médicos hipocráticos en el siglo IV a.C., diagnosticar una enfermedad era conocer su sintomatología individualmente, sabiendo distinguir entre ella y sus similares. El diagnóstico de las enfermedades ha sido posible por las habilidades que poseen especialistas en medicina humana, y que a partir de sus conocimientos y de la exploración de los pacientes son capaces de saber si una persona tiene cierta enfermedad (Entralgo, 1981). Esto es, el médico realiza una serie de preguntas a la persona y con base en las respuestas el especialista puede concluir, con cierta certeza, si dicho individuo se encuentra enfermo.

A través del tiempo el diagnóstico de enfermedades ha sufrido importantes cambios, especialmente en la segunda revolución industrial comprendida de 1850 a 1914, esto debido a la investigación en numerosos campos de la ciencia, y fue justo en 1895 donde por obra de la casualidad durante estudios con fenómenos eléctricos se descubrieron los rayos X. Este descubrimiento dio lugar a la búsqueda de nuevas tecnologías con el fin de diagnosticar con ayuda de imágenes. Después de la segunda guerra mundial llegaron nuevos avances como ecografía, tomografía, resonancia magnética, radio-protección sistemas de visualización y el más importante en 1951, la primera computadora comercial (Gálvez, 2013).

En 1956 surge la inteligencia artificial y comienzan a haber avances importantes en la computación. En 1986 se crean las redes neuronales recurrentes y es posterior a 1993 donde se comienza con una evolución constante en todo lo que a inteligencia artificial se refiere, y siendo uno de los momentos más importantes cuando Fei-Fei Li lanzó ImageNet en 2009, una base de datos basada en imágenes, que resultó en la reciente revolución del aprendizaje profundo (Abeliuk y Gutiérrez, 2021).

A partir de los avances que se han tenido en las últimas décadas en cuanto a inteligencia artificial, los médicos pueden contar con herramientas que los apoyan para el diagnóstico médico. La inteligencia artificial en el ámbito médico tiene como tarea diagnosticar o bien predecir enfermedades, a través del procesamiento de datos históricos que le permiten aprender las características comunes de cierta enfermedad (Díaz, 2019).

En la actualidad lo que se busca lograr con la inteligencia artificial en el ámbito de la salud es reducir el error al momento de diagnosticar enfermedades (Díaz, 2019).

1.2. Planteamiento del problema

1.2.1. Definición del problema

México ocupa el segundo lugar a nivel mundial en muertes por la diabetes mellitus tipo 2 (Rodríguez-Robles y Mayek-Perez, 2022) y según datos estadísticos presentados por la Secretaria de Salud en el IV Congreso Internacional y X Nacional de la Asociación de Enfermería Comunitaria (AEC), llevado a cabo en 2022, Tabasco ocupa el primer lugar nacional en detección de diabetes mellitus tipo 2 (DM2).

Cada año se registran 3,400 casos nuevos, haciendo énfasis en que sólo el 40 % de las personas enfermas con DM2 cuentan con un diagnóstico médico y están bajo control de un especialista de la salud. El resto lo desconoce y por lo tanto no tiene un tratamiento (Lázaro et al., 2022). Derivado de esto, la DM2 trae consigo complicaciones más allá de la propia enfermedad y siendo éstas las mismas que llevan al paciente a un deterioro significativo o la muerte, en donde la cardiopatía isquémica crónica (Hodelín Maynard et al., 2018), es una de las principales complicaciones que

afecta al paciente con DM2.

Aunado a los problemas de la misma diabetes y las complicaciones coronarias, el costo tanto para el paciente como para los sistemas de salud públicos son elevados. Se estima que las personas enfermas con DM2 gastan al menos el 200 % más de dinero en cuidados médicos, que quienes no tienen la enfermedad (Mendoza Romo et al., 2018), ya que se ataca el problema de diabetes y la complicación a la vez, eso sin asegurar que el paciente tendrá un estilo de vida dentro del mínimo confort. Es por ello que es necesario llevar el control del paciente con DM2 y desde un principio (dados ciertos factores), diagnosticar de manera temprana la posible complicación futura.

Existen casos de éxito donde mediante el análisis de ciertos parámetros se han conseguido buenos resultados en el diagnóstico oportuno de la DM2. Pero a decir de los especialistas médicos, los factores cambian respecto a la localidad, ya que también depende del área geográfica y estilo de vida. Un índice de desarrollo humano bajo se ha asociado con un alza de muertes por enfermedades crónicas, lo que refleja la presencia de desigualdad en el acceso a los servicios de salud, la calidad, la infraestructura y la cobertura (Mendoza Romo et al., 2018). Esto hace un cambio en el comportamiento de la enfermedad y en los factores claves que determinan una complicación, en este caso coronaria.

A partir de lo analizado hasta este momento se puede decir que el diagnóstico oportuno de la cardiopatía isquémica crónica representa un problema social que debe ser atendido, ya que cuando no se realiza un diagnóstico de forma preventiva se presentan complicaciones que deterioran significativamente la calidad de vida del paciente y pueden causarle la muerte; además, los costos económicos del tratamiento son grandes para el afectado, así como para los sistemas públicos de salud.

Hasta el momento existen trabajos con redes neuronales artificiales donde se ha podido detectar con una precisión del 87.3 %, si una persona tiene diabetes (El Jerjawi y Abu-Naser, 2018), además de que con una red neuronal es posible clasificar entre personas con DM2 y un grupo sano, con una precisión de predicción excelente (Siptroth et al., 2023). Estos antecedentes muestran que las redes neuronales pueden predecir los factores de riesgo de la enfermedad.

En 2020 un estudio basado en redes neuronales convolucionales y realizado en Estados Unidos

a pacientes mexicanos logró clasificar tres distintas cardiopatías (Baccouche et al., 2020). En otro estudio que empleó visión Transformer(ViT) se pudo clasificar cardiopatías a través de los sonidos del corazón, aunque con datos sintéticos (Kinha et al., 2023). También los avances en las redes neuronales Transformer han permitido que en nuevos estudios se pueda detectar cardiopatías congénitas a partir de grabaciones de fono cardiograma (Alkhodari et al., 2022). En 2022, haciendo uso de redes neuronales Transformer, se logró detectar soplos cardíacos a través del filtrado de sonidos del corazón (Fan et al., 2022).

Todos los avances mencionados con respecto al estudio de cardiopatías, en un principio con CNN y luego con redes neuronales Transformer, están basados en datos de pacientes de otras latitudes o con datos artificiales. Por tanto, a pesar de que los resultados son favorables en la mayoría de los casos, es pertinente realizar estudios más localizados.

A nuestro conocimiento no existen investigaciones donde se esté trabajando en la detección de cardiopatía isquémica crónica en pacientes con DM2 con ayuda de redes neuronales Transformer, es por esto que es necesario realizar un análisis local, con estudios de pacientes mexicanos y determinar el modelo de red neuronal artificial (atributos de entrada, número de capas, cantidad de neuronas por capa, función de activación y parámetros en general) que nos permita diagnosticar la enfermedad coronaria, como una complicación de la DM2, con una precisión de al menos el 80 %. En este estudio se realiza un análisis del *dataset* `hk_database` (Garcia, 2023) de la colección DiabetIA. La colección DiabetIA es resultado de un proyecto de Ciencia de Frontera financiado por el CONAHCYT (actualmente SEHCITI) y contiene datos reales de pacientes mexicanos.

1.2.2. Delimitación de la investigación

Alcances

- El modelo implementado únicamente considera el diagnóstico, sin llegar a la prevención o tratamiento.
- La investigación se enfoca en encontrar el patrón de diagnóstico para la cardiopatía isquémica crónica, sin considerar otras enfermedades incluidas en el *dataset*.

- Esta investigación se centra sólo en implementar el modelo de red neuronal que pueda diagnosticar la cardiopatía isquémica crónica, sin considerar la implementación en escenarios reales.

Limitaciones

- La calidad de los resultados está en función de la calidad de los datos disponibles en el *dataset* `hk_database`.
- La generalización de los resultados está supeditada por las características y cantidad de datos disponibles en el *dataset* `hk_database`.
- El diseño experimental comprende los hiperparámetros más comunes utilizados en las redes neuronales.
- Esta investigación se centra sólo en utilizar el *dataset* `hk_database`. para entrenar el modelo que contiene la información clínica de pacientes mexicanos con DM2.

1.3. Preguntas de investigación e hipótesis

1.3.1. Preguntas de investigación

¿Cuál es la estructura de red neuronal artificial en cuanto a capas, neuronas por capa y funciones de activación que diagnostique la cardiopatía isquémica crónica con al menos el 80 % de exactitud?

¿Cuáles son las métricas de calidad más adecuadas para evaluar la precisión de la red neuronal artificial Transformer?

1.3.2. Hipótesis

Una red neuronal artificial Transformer puede diagnosticar la cardiopatía isquémica crónica en enfermos con diabetes mellitus tipo 2 con al menos un 80 % de precisión.

1.4. Objetivo general

Implementar un modelo de red neuronal artificial Transformer para diagnosticar la cardiopatía isquémica crónica en pacientes con diabetes mellitus tipo 2.

1.5. Objetivos específicos

- Efectuar un análisis exploratorio del *dataset* `hk_database`.
- Desarrollar un modelo de red neuronal Transformer que diagnostique la cardiopatía isquémica crónica.
- Evaluar el rendimiento del modelo propuesto.
- Efectuar las adecuaciones necesarias para alcanzar la mayor efectividad posible, siendo al menos de 80 %.
- Analizar los resultados del modelo implementado de red neuronal Transformer con respecto a otros modelos de aprendizaje automático.

1.6. Justificación

Una persona con cardiopatía isquémica crónica puede pasar parte de su vida con el padecimiento, incluso sin saberlo, y en el peor de los casos puede sufrir un infarto. Para una persona con cardiopatía isquémica crónica, el costo de tratamiento puede ser muy elevado, eso en dado caso de saber que ya la padece. De tener sospechas que tiene la enfermedad dependerá de los estudios que su médico le indique, los cuales en su mayoría son costosos (desde un electrocardiograma hasta un cateterismo cardíaco), para así el medico decir si la padece o no, esto sin dejar de lado que, para estos casos, son personas con diabetes.

Debido a la transformación digital, gran parte de nuestros datos están ahora digitalizados y dentro de esos datos se encuentran nuestros datos médicos. Gracias a esto, con el tratamiento correcto de ellos y con las tecnologías emergentes como las redes neuronales Transformer, es posible detectar con precisión las enfermedades . Es por esto que, de encontrar el modelo de red neuronal

artificial Transformer adecuado que nos indique si una persona padece cardiopatía isquémica crónica, el paciente y sus familiares principalmente, tendrían ahorros considerables, esto se debe a que podrían tratar la enfermedad de manera adecuada y evitar complicaciones mayores, como un infarto.

1.7. Metodología utilizada

En este apartado se presenta una breve descripción de la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) utilizada para lograr los objetivos del presente proyecto.

- Identificar los objetivos y requisitos de la investigación. Para convertir este conocimiento en un planteamiento de problema y un plan preliminar diseñado para lograr los objetivos.
- Comprender los datos. Iniciando por la recopilación u obtención de ellos y familiarizarse con los mismos, a fin de identificar problemas de calidad o detectar subconjuntos de datos.
- Preparar los datos. Cubre las actividades que se realizan para conseguir los datos que se utilizan en el modelo, a partir de los datos iniciales sin procesar, en esta fase se realiza un análisis exploratorio de datos (EDA, por las siglas en Inglés de *Exploratory Data Analysis*) para garantizar el manejo correcto de los datos.
- Implementar el modelo de red neuronal Transformer y calibrar parámetros a los valores óptimos. En esta fase se entrena el modelo.
- Evaluar el modelo de red neuronal Transformer contra otros modelos de aprendizaje automático.
- Presentar el modelo una vez validada su funcionalidad.

Capítulo 2

Marco teórico

2.1. Conceptos y teorías fundamentales de la investigación

En esta sección se efectúa una revisión bibliográfica de los conceptos computacionales y médicos fundamentales, que son base para el análisis conceptual de esta investigación.

2.1.1. Red neuronal artificial

Una red neuronal artificial (ANN, por las siglas en inglés de *Artificial Neural Network*) es un modelo computacional que está inspirado en las características y operación del cerebro humano, y realiza un proceso de tipo *machine learning*. Una de las características de las ANNs es aprender de sus errores y mejorar continuamente (AWS, 2023; Martin y Sanz Molina, 2001).

La sigla ANN es utilizada para referirse al conjunto general de redes neuronales artificiales, ya que dentro de las ANNs existen diferentes arquitecturas y tipos de redes neuronales artificiales. Cada una tiene características y aplicaciones específicas, pero todas están bajo el concepto general de las redes neuronales artificiales.

2.1.2. Red neuronal artificial Transformer

Un modelo Transformer es una red neuronal artificial que se caracteriza por su capacidad para manejar secuencias de datos de manera efectiva y capturar relaciones de largo alcance entre

elementos de la secuencia, una de sus características principales es que fue desarrollada para el procesamiento de lenguaje natural (Nerella et al., 2023).

Las redes neuronales Transformer inicialmente surgieron para el procesamiento de lenguaje natural y su característica especial es el módulo de atención con el que cuenta, el cual le permite aprender del contexto, y por lo tanto, el significado existente mediante el seguimiento de relaciones entre datos secuenciales. Debido a esta característica significativa, la red neuronal Transformer y su capacidad de aprender ya está siendo utilizada en otras áreas.

2.1.3. TabTransformer

El TabTransformer es una adaptación específica del modelo Transformer, TabTransformer fue diseñado para el procesamiento de datos tabulares. A diferencia del Transformer original, que fue diseñado para procesar secuencias de texto, el TabTransformer está optimizado para manejar datos estructurados con características tanto categóricas como numéricas (X. Huang et al., 2020).

Sus características principales incluyen:

- **Procesamiento dual:** Maneja simultáneamente variables categóricas mediante *embeddings* y variables continuas a través de normalización especializada.
- **Mecanismo de atención adaptado:** Modifica el mecanismo de atención tradicional para capturar relaciones entre diferentes características tabulares, permitiendo que el modelo aprenda patrones complejos en los datos.
- ***Column embeddings*:** Implementa *embeddings* específicos para cada columna categórica, permitiendo que el modelo aprenda representaciones únicas para cada tipo de característica.
- ***Feature normalization*:** Aplica normalización especializada para características continuas, facilitando el entrenamiento y mejorando la convergencia del modelo.

2.1.4. Métricas de evaluación en clasificación médica

En el contexto del diagnóstico médico, la selección y comprensión de las métricas de evaluación es crucial, ya que diferentes tipos de errores pueden tener distintas implicaciones clínicas (Hossin y Sulaiman, 2015):

- **Exactitud (*Accuracy*):** Mide la proporción total de predicciones correctas. En el contexto médico, aunque es importante, no debe ser la única métrica considerada, ya que no distingue entre diferentes tipos de errores (Sokolova y Lapalme, 2009):

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

  donde VP son Verdaderos Positivos, VN son Verdaderos Negativos, FP son Falsos Positivos y FN son Falsos Negativos.

- **Precisión (*Precision*):** Mide la proporción de verdaderos positivos entre todos los casos clasificados como positivos:

$$Precision = \frac{VP}{VP + FP} \tag{2.2}$$

  Esta métrica es especialmente importante cuando se quiere minimizar los falsos positivos.

- **Sensibilidad (*Recall*):** Especialmente crucial en diagnósticos médicos, mide la capacidad del modelo para identificar correctamente los casos positivos (Powers, 2020):

$$Recall = \frac{VP}{VP + FN} \tag{2.3}$$

  En diagnósticos médicos, un alto *recall* es prioritario cuando no detectar una enfermedad (falso negativo) puede resultar en tratamientos tardíos o ausentes, lo que podría comprometer la salud del paciente.

- ***F1-Score*:** Proporciona un balance entre precisión y *recall*, especialmente útil en conjuntos de datos desbalanceados donde hay una clase minoritaria importante (Chicco y Jurman, 2020):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

- **AUC-ROC:** (*Area Under the Curve - Receiver Operating Characteristic*) Evalúa la capacidad discriminativa del modelo a través de diferentes umbrales de clasificación (Fawcett, 2006):
  - Un valor de 1.0 representa una clasificación perfecta.
  - Un valor de 0.5 representa una clasificación aleatoria.
  - Es especialmente útil en diagnósticos médicos donde el umbral de decisión puede ajustarse según el contexto clínico.

2.1.5. Técnicas de validación y optimización

El desarrollo de modelos robustos para aplicaciones médicas requiere técnicas específicas de validación y optimización:

- **Validación cruzada:** Metodología que permite evaluar la capacidad de generalización del modelo (Kohavi, 1995):
  - Divide los datos en k subconjuntos (*folds*).
  - Entrena el modelo k veces, usando cada vez un subconjunto diferente como validación.
  - Proporciona una estimación más confiable del rendimiento real del modelo.
  - Especialmente importante en contextos médicos donde los datos pueden ser limitados.
- ***Early stopping*:** Técnica para prevenir el sobreajuste mediante la detención temprana del entrenamiento (Prechelt, 2002):
  - Monitoreo continuo del rendimiento en el conjunto de validación.
  - Detención del entrenamiento cuando el rendimiento deja de mejorar.
  - Definición de un período de "paciencia" para evitar detenciones prematuras.
  - Preservación de los mejores pesos del modelo durante el entrenamiento.
- **Algoritmos genéticos para optimización de hiperparámetros:** Método de búsqueda inspirado en la evolución natural para encontrar la configuración óptima de parámetros (Han y Xiao, 2022):

- **Población:** Conjunto de posibles configuraciones de hiperparámetros que representan soluciones candidatas al problema.
- ***Fitness*:** Evaluación del rendimiento de cada configuración según criterios predefinidos.
- **Selección:** Elección de las mejores configuraciones para reproducción, basada en el principio de "supervivencia del más apto".
- ***Crossover* adaptativo:** Combinación de configuraciones exitosas con probabilidades que se ajustan según el valor de aptitud (*fitness*) de los individuos.
- **Mutación adaptativa:** Introducción de variaciones aleatorias con tasas que se adaptan dinámicamente durante el proceso de optimización, evitando la convergencia prematura a óptimos locales.

2.1.6. La diabetes mellitus tipo 2

La diabetes mellitus tipo 2 (DM2) es una condición metabólica que se caracteriza por la resistencia a la insulina o por la incapacidad del cuerpo para utilizarla eficazmente. En este trastorno, aunque el páncreas produce insulina, las células muestran una disminución en su respuesta a esta hormona, esto resulta en un desequilibrio en los niveles de glucosa en la sangre, ya que el cuerpo no puede utilizar adecuadamente la glucosa para obtener energía, lo que conduce a niveles elevados de azúcar en la sangre (hiperglucemia). Una característica notable de la diabetes tipo 2 es su asociación con factores de riesgo como la obesidad, la falta de actividad física, la alimentación poco saludable y la predisposición genética. Estos factores pueden influir en el desarrollo y la progresión de la enfermedad (MedlinePlus, 2023).

La enfermedad DM2 está ligada al estilo de vida de quien la desarrolla, ya que es más común en personas que tienen sobrepeso u obesidad, así como aquellos que no practican actividad física, y normalmente se desarrolla más en personas adultas que en niños o adolescentes. La DM2 se caracteriza porque el cuerpo no puede utilizar de manera adecuada la insulina y se elevan los niveles de glucosa en la sangre. Si la DM2 no se controla adecuadamente, las afectaciones antes mencionadas pueden ocasionar complicaciones a largo plazo, tales como daño a los vasos

sanguíneos, nervios, ojos y otros órganos. El manejo de la DM2 incluye cambios en el estilo de vida, como dieta equilibrada, ejercicio regular y, en algunos casos, medicamentos para controlar los niveles de azúcar en la sangre.

2.1.7. Cardiopatía isquémica

La cardiopatía isquémica es un padecimiento que se caracteriza por el estrechamiento del diámetro interno de las arterias que suministran sangre al corazón, y esa reducción es causada por una acumulación de placa, lo cual aumenta drásticamente la posibilidad de sufrir un infarto al miocardio. Una característica de este padecimiento es tener la enfermedad y que no presente síntoma alguno (FEC, 2024a).

La cardiopatía isquémica es una enfermedad que se desarrolla con el paso del tiempo, y se caracteriza por el proceso lento de formación de colágeno y acumulación de lípidos (grasas) y células inflamatorias (linfocitos), lo cual puede desencadenar dolores en el pecho o un infarto.

2.2. Marco tecnológico

En esta lista se describen el *software* y *hardware* que se ha utilizado para la elaboración de este trabajo.

- Lenguaje de Programación: Python 3.12.3. Python es ampliamente utilizado en el desarrollo de modelos de aprendizaje profundo debido a su sintaxis sencilla y la extensa biblioteca de herramientas disponibles.
- Entorno de Desarrollo: R versión 4.4.0. R es conocido por sus capacidades estadísticas y gráficas, siendo muy útil para el análisis de datos.
- Entorno de Desarrollo Integrado (IDE, por las siglas en inglés *Integrated Development Environment*): RStudio *Build* 748. RStudio proporciona una interfaz amigable y poderosa para desarrollar en R, facilitando el manejo de proyectos y la visualización de datos.
- PyTorch: Este *framework* proporciona herramientas y bibliotecas avanzadas para el desarrollo y entrenamiento de modelos de aprendizaje profundo.

- *Drivers* y Bibliotecas: CUDA y cuDNN, junto con los controladores adecuados para la Unidad de Procesamiento Gráfico(GPU, por las siglas en inglés de *Graphics Processing Unit*), son esenciales para optimizar el rendimiento computacional y acelerar el entrenamiento de redes neuronales.
- Procesador (CPU, por las siglas en inglés de *Central Processing Unit*): Intel Core i7-13700K. Este procesador ofrece un rendimiento sólido para manejar tareas intensivas, ideal para el desarrollo y entrenamiento de modelos de aprendizaje profundo.
- Unidad de Procesamiento Gráfico (GPU): NVIDIA RTX 4060 Ti. Las GPUs de NVIDIA son altamente recomendadas debido a su compatibilidad con CUDA y cuDNN, herramientas esenciales para la optimización y aceleración del entrenamiento de redes neuronales.
- Memoria RAM: 32 GB. Una cantidad significativa de memoria es crucial para gestionar grandes volúmenes de datos y realizar múltiples operaciones simultáneamente de manera eficiente.
- Almacenamiento: Unidad de Estado Sólido (SSD, por las siglas en inglés de *Solid State Drive*) de 1TB. Las SSDs son preferidas por su alta velocidad de lectura y escritura, lo que mejora significativamente el rendimiento general del sistema y reduce los tiempos de carga de datos.

Cada uno de estos componentes ha sido seleccionado meticulosamente para asegurar un entorno de desarrollo robusto y eficiente, adecuado para las tareas complejas de procesamiento y análisis de datos necesario en este proyecto.

2.3. Literatura relacionada

En la Tabla 2.1 se muestra de manera resumida y en orden cronológico, algunos trabajos basados en la arquitectura de red neuronal artificial Transformer.

Tabla 2.1

Tabla en orden cronológico de autores y contribuciones/descubrimientos.

Año	Autor(es)	Contribución / Descubrimiento
2020	Kexin Huang, Jaan Altosaar y Rajesh Ranganath (K. Huang et al., 2019).	Modelo Transformer para predecir la readmisión hospitalaria de pacientes a partir de la fecha del alta médica.
2021	Ivan Matas González (Matas González, 2021).	Se demuestra que ViT es mejor al momento de clasificar imágenes que CNN.
2021	Alireza Roshanzamir, Hamid Aghajan y Mahdieh Soleymani Baghshah (Roshanzamir et al., 2021).	Modelos de aprendizaje profundo con base en Transformer y PLN para la evaluación temprana del Alzheimer a partir del test de descripción de imágenes.
2021	Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li y Aichi Chien (Monajatipoor et al., 2022).	Modelos Transformer de visión y lenguaje para mejorar el diagnóstico de enfermedades a través de imágenes.
2022	Mohanad Alkhodari, Syafiq Kamarul Azman y Leontios J. Hadjileontiadis (Alkhodari et al., 2022).	Modelo Transformer para detectar enfermedades congénitas del corazón por medio de grabaciones de PCG.
2022	Pengfei Fan, Yucheng Shu y Yiming Han (Fan et al., 2022).	Modelo de red neuronal Transformer para la detección de soplos cardíacos.
2023	Oguzhan Katar y Ozal Yildirim (Katar y Yildirim, 2023).	Modelo de visión Transformer (ViT) para la detección automática de glóbulos blancos a partir de películas sanguíneas.
2023	Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz y Kazem Rahimi (Y. Li et al., 2022)	Modelo basado en Transformer para predicción de riesgos de diversas enfermedades.
2023	Jun Li, Junyu Chen, Yucheng Tang, Ce Wang y Bennett A. Landman (J. Li et al., 2023).	Modelo basado en Transformer mejora la visión por computadora de modelos basados en convolucionales.

Capítulo 3

Modelo de la red neuronal artificial Transformer

3.1. *Dataset*

En este capítulo se presenta el *dataset* `hk_database`, el cual forma parte del repositorio "DiabetIA" construido a partir de un proyecto financiado por el CONAHCYT (actualmente SEHCITI) y ha sido utilizado como base para esta investigación. Este *dataset* contiene el registro de pacientes mexicanos con DMT2 y con complicaciones crónicas. El conjunto de datos fue extraído de la información contenida en el Sistema de Informática de Medicina Familiar (SIMF) del Instituto Mexicano del Seguro Social (IMSS), en el estado de Michoacán, México. La creación de este *dataset* surgió como parte de una convocatoria emitida en junio de 2019 dentro de los Programas Nacionales Estratégicos (PRONACES), bajo el título del proyecto "Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2".

El *dataset* contiene un total de 520 variables divididas en seis grupos, además de 477,036 observaciones. Este conjunto de datos comprende información recopilada desde el año 2005, como el primer ingreso, hasta el 2020, considerando la última actualización realizada el 1 de septiembre de 2021. Este conjunto de datos será utilizado para analizar la relación entre la diabetes mellitus tipo 2 y la cardiopatía isquémica crónica.

Los seis grupos en los que se dividen las variables del *dataset* son los siguientes: *Demographics, Medical Diagnosis, Measurements, Laboratories, Drug* y *Predictions*. Cabe destacar que los datos están sin procesar, lo que requiere un preprocesamiento previo para su análisis.

3.2. Análisis exploratorio de los datos

El análisis exploratorio de datos es una etapa crucial en el preprocesamiento, ya que transforma los datos en un formato adecuado para ser procesados de manera eficiente en tareas de minería de datos, aprendizaje automático y otras áreas de la ciencia de datos. El preprocesamiento asegura que los datos estén en óptimas condiciones para obtener resultados precisos y robustos durante el desarrollo del modelo.

3.2.1. Pre-selección de variables

El *dataset* `hk_database` incluye un diccionario de datos claramente definido, dividido en las siguientes cuatro columnas: *group, column, type* y *description*. Este diccionario permite identificar rápidamente la naturaleza de las variables y su utilidad en el análisis. A continuación, se presenta como ejemplo el primer grupo de variables en la Tabla 3.1, el cual sirvió como referencia inicial para analizar el tipo de datos y la información contenida en cada variable.

Tabla 3.1

Grupo Demographics del diccionario de datos.

Group	column	Type	Description
Demographics (17)	id	string	Unique ID for patient for window
	cx_curp	string	Patient unique ID
	birthdate	date	Date of birth
	cs_sex	binary	Sex
	careunit	string	Patient's careunit
	first_cx	date	First historic visit date registered
	last_cx	date	Last historic visit date registered
	count_cx_w	integer	Number of visits (por ventana)
	years_cx	integer	Years of follow up
	window	integer	A window represents the summarized data of two years defined by x_start and x_end dates
	x_start	date	Start date of time period for values
	x_end	date	End date of time period for values
	y_start	date	Start date of time period for diagnosis predictions
	y_end	date	End date of time period for diagnosis predictions
	age_at_wx	integer	Patient age at the end of the time period for values (x_end)
	dx_year_e11	integer	Diabetes mellitus type 2 diagnosis year
	dx_age_e11	integer	Age at Diabetes mellitus type 2 diagnosis

A partir del análisis del diccionario de datos, se identificó la necesidad de eliminar ciertas variables categóricas que no aportaban información relevante para el estudio. Además, se decidió excluir por completo el grupo *Predictions*, ya que este hacía referencia a predicciones generadas en la investigación original de la cual se derivó el conjunto de datos.

La Tabla 3.2 presenta los dos grupos donde se realizó esta primera selección de variables. Las variables sombreadas corresponden a aquellas que fueron descartadas durante este proceso. Como resultado, el *dataset* quedó conformado por 498 variables, reduciendo las 520 iniciales.

Tabla 3.2

Tabla que muestra las variables eliminadas en la pre-selección.

Group	column	Type	Description
Demographics (17)	id	string	Unique ID for patient for window
	cx_curp	string	Patient unique ID
	birthdate	date	Date of birth
	cs_sex	binary	Sex
	careunit	string	Patient's careunit
	first_cx	date	First historic visit date registered
	last_cx	date	Last historic visit date registered
	count_cx_w	integer	Number of visits (por ventana)
	years_cx	integer	Years of follow up
	window	integer	Number of patient window (time period defined). A window represents the summarized data of two years defined by x_start and x_end dates
	x_start	date	Start date of time period for values
	x_end	date	End date of time period for values
	y_start	date	Start date of time period for diagnosis predictions
	y_end	date	End data of time period for diagnosis predictions
	age_at_wx	integer	Patient age at the end of the time period for values (x_end)
	dx_year_e11	integer	Diabetes mellitus type 2 diagnosis year
	dx_age_e11	integer	Age at Diabetes mellitus type 2 diagnosis
Predictions (11)	e11	binary	Type 2 diabetes mellitus (ICD E11) diagnosis in following year
	e110	binary	Type 2 diabetes mellitus with coma (ICD E11.0) diagnosis in following year
	e111	binary	Type 2 diabetes mellitus with ketoacidosis (ICD E11.1) diagnosis in following year
	e112	binary	Type 2 diabetes mellitus with kidney complications (ICD E11.2) diagnosis in following year
	e113	binary	Type 2 diabetes mellitus with ophthalmic complications (ICD E11.3) diagnosis in following year
	e114	binary	Type 2 diabetes mellitus with neurological complications (ICD E11.4) diagnosis in following year
	e115	binary	Type 2 diabetes mellitus with peripheral circulatory complications (ICD E11.5) diagnosis in following year
	e116	binary	Type 2 diabetes mellitus with other specified complications (ICD E11.6) diagnosis in following year
	e117	binary	Type 2 diabetes mellitus with multiple complications (ICD E11.7) diagnosis in following year
	e118	binary	Type 2 diabetes mellitus with unspecified complications (ICD E11.8) diagnosis in following year
	e119	binary	Type 2 diabetes mellitus without mention of complication (ICD E11.9) diagnosis in following year

Nota. Elaboración propia con base en el diccionario de datos del conjunto.

3.2.2. Datos nulos

En un conjunto de datos es común que puedan existir datos nulos, esto es, columnas y renglones que no tienen información. Esta situación puede atenderse de dos maneras. La primera de ellas es rellenar (imputar) con valores estadísticos u otras técnicas, aquellos datos faltantes, siempre y cuando cumplan con el criterio establecido en la literatura para estos casos. Lo cual establece que la cantidad de datos imputados no debe ser mayor al 25 % de la cantidad total de datos en la columna (variable) o ese mismo porcentaje respecto al total de renglones (observaciones) (Galván y Medina, 2007). La segunda opción sería eliminar por completo la columna o los renglones. Esto sería para el caso que la cantidad de datos faltantes supere el porcentaje determinado para procesos de imputación.

3.2.2.1. Datos nulos por observaciones

Aunque el *dataset* fue reducido en cuanto al número de variables en la etapa de pre-selección (véase Tabla 3.2), persistía una desproporción significativa entre las observaciones de las clases “personas diabéticas con cardiopatía isquémica crónica” y “personas diabéticas sin cardiopatía isquémica crónica”, con 8,364 observaciones frente a 112,407 respectivamente.

Dado lo anterior, se realizó un análisis de los datos nulos por observaciones, enfocándose en la clase “personas diabéticas sin cardiopatía isquémica crónica”, debido a que esta clase contaba con 112,407 observaciones. Este análisis tuvo como objetivo conservar la mayor cantidad de datos completos en cada observación.

La Tabla 3.3 muestra el rango de datos nulos al recorrer cada renglón, encontrándose valores faltantes entre 17 y 178, lo que corresponde aproximadamente a un 3.5 % y 35.8 % de datos nulos sobre el total de variables, respectivamente. Para abordar esta situación, se decidió filtrar las observaciones con la menor cantidad de datos faltantes, estableciendo un máximo de 60 valores nulos por observación, equivalente a un 12 % de valores faltantes. Este criterio permitió obtener un total de 8,585 observaciones para su posterior análisis.

Tabla 3.3

Cantidad de nulos por observación para "personas diabéticas sin cardiopatía isquémica crónica".

Clase 2 (Sin cardiopatía)	
Número de observación	**Nulos**
59534	17
13519	19
31765	22
63423	22
94122	22
25697	23
...	...
112297	178
112298	178
112392	178
112393	178

Nota. Se muestran los valores extremos de la distribución de datos nulos.

Aunque en la tabla no se presenta la totalidad de las observaciones con datos nulos, se tiene que 103,822 observaciones fueron descartadas únicamente mediante el proceso de filtrado.

3.2.2.2. Datos nulos por variables

Dentro del análisis de los datos nulos por variables, se observó que algunas presentaban valores faltantes en varias observaciones. La Tabla 3.4 muestra un desglose de la cantidad de datos nulos identificados en las diferentes variables del *dataset*.

Tabla 3.4

Cantidad de nulos por variable para "personas diabéticas con cardiopatía isquémica crónica" y "personas diabéticas sin cardiopatía isquémica crónica".

Clase 1 (Con cardiopatía)			Clase 2 (Sin cardiopatía)		
Col	**Column**	**Nulos**	**Col**	**Column**	**Nulos**
333	fn_urine_culture_std	8364	333	fn_urine_culture_std	103822
328	fn_urine_culture_mean	8363	328	fn_urine_culture_mean	103796
329	fn_urine_culture_median	8363	329	fn_urine_culture_median	103796
330	fn_urine_culture_max	8363	330	fn_urine_culture_max	103796
331	fn_urine_culture_min	8363	331	fn_urine_culture_min	103796
334	fn_urine_culture_slope	8363	334	fn_urine_culture_slope	103796
298	fn_urea_std	8309	298	fn_urea_std	103614
326	fn_ego_density_std	8308	326	fn_ego_density_std	103359
291	fn_ego_std	8306	291	fn_ego_std	103356
319	fn_creatinine_std	8274	312	fn_aurico_std	103097
312	fn_aurico_std	8263	319	fn_creatinine_std	103064
242	fn_capillary_glucose_std	8234	293	fn_urea_mean	102591
293	fn_urea_mean	8185	294	fn_urea_median	102591
294	fn_urea_median	8185	295	fn_urea_max	102591
295	fn_urea_max	8185	296	fn_urea_min	102591
	...			...	
213	in_resp_freq_count	143	213	in_resp_freq_count	5400
220	fn_temperature_count	143	220	fn_temperature_count	5400
227	in_glucose_count	143	227	in_glucose_count	5400
5	count_cx_w	136	5	count_cx_w	4832
1	x	0	1	x	0
2	id	0	2	id	0

Nota. Se presentan las variables con mayor cantidad de valores nulos, ordenadas de mayor a menor.

Aunque en la tabla anterior no se presenta la cantidad completa de variables con datos nulos para cada una de las clases, se identificó un total de 178 variables con valores faltantes en ambas clases. Además, se observó que tanto en la clase "personas diabéticas con cardiopatía isquémica crónica" como en la clase "personas diabéticas sin cardiopatía isquémica crónica", existen variables donde la totalidad o la mayoría de los datos son nulos.

3.2.3. *Sub-sampling*

Después del filtrado por observaciones se pudo notar que las clases seguían un poco desbalanceadas. Por esta razón, fue necesario balancear los *subsets* antes de proceder con su unión. Se

decidió emplear el método de *Sub-sampling*, una técnica que consiste en seleccionar aleatoriamente muestras de la clase mayoritaria y ajustarlas al tamaño de la clase minoritaria (Espinar Lara, 2018; Ignacio, 2019).

Para garantizar la reproducibilidad del proceso, se utilizó una semilla con valor 123. Además, la cantidad de observaciones en el *subset* resultante se fijó en 8,364, igualando al tamaño de la clase minoritaria.

3.2.4. Eliminación de columnas con valores nulos

Una vez extraída la mayor cantidad de datos por observaciones y generado el *sub-sampling* para la clase mayoritaria, se procedió a filtrar las variables con menor cantidad de datos nulos. Este proceso tuvo como objetivo identificar las variables con la mayor proporción de datos completos, realizándose de forma independiente para ambas clases y para cada uno de los grupos de variables.

Se estableció un umbral máximo de datos nulos de 2,091, lo que equivale al 25 % de valores faltantes por variable, considerando un total de 8,364 observaciones por clase. Este criterio fue tomado como referencia para conservar únicamente las variables más completas en cada grupo.

3.2.4.1. *Demographics*

La Tabla 3.5 muestra la cantidad de datos nulos después del filtrado en el grupo *Demographics*. Como se observa, solo la clase 1 (personas diabéticas con cardiopatía isquémica crónica) presenta 136 datos nulos en una variable, mientras que la clase 2 (personas diabéticas sin cardiopatía isquémica crónica) no contiene valores faltantes.

Tabla 3.5

Cantidad de nulos en la clase 1 y clase 2 para el grupo demographic.

Clase 1 (Con cardiopatía)		Clase 2 (Sin cardiopatía)	
Column	**Nulos**	**Column**	**Nulos**
id	0	id	0
cx_curp	0	cx_curp	0
cs_sex	0	cs_sex	0
count_cx_w	136	count_cx_w	0
age_at_wx	0	age_at_wx	0
dx_age_e11	0	dx_age_e11	0

Nota. Se muestra el número de valores nulos en cada variable del grupo *demographic* para ambas clases de pacientes.

3.2.4.2. *Medical Diagnosis*

La Tabla 3.6 presenta la cantidad de datos nulos después del proceso de filtrado en el grupo *Medical Diagnosis*. Como se observa, este grupo no contiene datos nulos en ninguna de las 166 variables analizadas para ambas clases.

Tabla 3.6

Cantidad de nulos en la clase 1 y clase 2 para el grupo medical diagnosis.

	Clase 1 (Con cardiopatía)		Clase 2 (Sin cardiopatía)	
	Column	**Nulos**	**Column**	**Nulos**
1	diabetes_mellitus_type_2	0	diabetes_mellitus_type_2	0
2	essential_primary_hypertension	0	essential_primary_hypertension	0
3	circulatory_system_diseases	0	circulatory_system_diseases	0
4	endocrine_nutritional_and_ metabolic_diseases	0	endocrine_nutritional_and_ metabolic_diseases	0
5	diseases_of_the_musculoskeletal_ system_and_connective_tissue	0	diseases_of_the_musculoskeletal_ system_and_connective_tissue	0
6	nervous_system_diseases	0	nervous_system_diseases	0
7	factors_influencing_health_status_ and_contact_with_health_services	0	factors_influencing_health_status_ and_contact_with_health_services	0
	⋮		⋮	
163	type_2_diabetes_with_renal_comp	0	type_2_diabetes_with_renal_comp	0
164	type_2_diabetes_with_neuro_comp	0	type_2_diabetes_with_neuro_comp	0
165	diabetes_mellitus_type_2_with_coma	0	diabetes_mellitus_type_2_with_coma	0
166	type_2_diabetes_with_ketoacidosis	0	type_2_diabetes_with_ketoacidosis	0

Nota. Se muestra que todas las variables del grupo *medical diagnosis* no contienen valores nulos en ninguna de las dos clases. Algunos nombres de variables han sido abreviados para mejorar la visualización.

3.2.4.3. *Measurements*

La Tabla 3.7 muestra las cantidades de datos nulos después del proceso de filtrado en el grupo *Measurements*. Como se observa, este grupo contiene datos nulos en ambas clases; sin embargo, las 49 variables iniciales permanecen presentes tras el filtrado.

Tabla 3.7

Cantidad de nulos en la clase 1 y clase 2 para el grupo measurements.

	Clase 1 (Con cardiopatía)		Clase 2 (Sin cardiopatía)	
	Column	**Nulos**	**Column**	**Nulos**
1	fn_weight_mean	143	fn_weight_mean	0
2	fn_weight_median	143	fn_weight_median	0
3	fn_weight_max	143	fn_weight_max	0
4	fn_weight_min	143	fn_weight_min	0
5	fn_weight_count	143	fn_weight_count	0
⋮	⋮	⋮	⋮	⋮
46	fn_temperature_min	176	fn_temperature_min	13
47	fn_temperature_count	143	fn_temperature_count	0
48	fn_temperature_std	249	fn_temperature_std	206
49	fn_temperature_slope	176	fn_temperature_slope	13

Nota. Se muestran los valores nulos para cada variable del grupo *measurements*, con diferencias notables entre las dos clases de pacientes.

3.2.4.4. *Laboratories*

La Tabla 3.8 muestra las cantidades de datos nulos en el grupo *Laboratories* para ambas clases. De las 128 variables originales en este grupo, únicamente 6 variables en la clase 1 (pacientes con cardiopatía) cumplieron con el umbral máximo establecido del 25 % de valores faltantes. Por otro lado, en la clase 2 (pacientes sin cardiopatía), 67 variables se encontraban dentro de este criterio, incluyendo las mismas 6 variables identificadas en la clase 1.

Para mantener la coherencia entre ambas clases y asegurar un análisis comparativo válido, se tomó como referencia la clase más restrictiva (clase 1) para el filtrado. En consecuencia, se conservaron únicamente las 6 variables que cumplían los criterios en ambas clases, resultando en la eliminación de 122 variables (128 - 6 = 122) del conjunto original de datos. Esta decisión priorizó la calidad de los datos sobre la cantidad, asegurando que solo se utilizaran variables con suficiente información confiable para el análisis posterior.

Tabla 3.8

Cantidad de nulos en la clase 1 y clase 2 para el grupo laboratories.

	Clase 1 (Con cardiopatía)			Clase 2 (Sin cardiopatía)	
	Column	**Nulos**		**Column**	**Nulos**
1	in_glucose_mean	1825	1	in_glucose_mean	1862
2	in_glucose_median	1825	2	in_glucose_median	1862
3	in_glucose_max	1825	3	in_glucose_max	1862
4	in_glucose_min	1825	...	...	...
5	in_glucose_count	143	66	leukocytes	0
6	in_glucose_slope	1825	67	others	0

Nota. Se presentan las variables del grupo *laboratories* con su respectiva cantidad de valores nulos por clase.

3.2.4.5. *Drug*

La Tabla 3.9 presenta las cantidades de datos nulos después del proceso de filtrado en el grupo *Drug*. Como se observa, este grupo no contiene datos nulos en ninguna de las dos clases.

Esto se debe a que, para este *dataset*, los datos nulos relacionados con medicamentos fueron tratados como medicamentos no prescritos a los pacientes, asignando valores de cero a las observaciones con datos faltantes (Tripp et al., 2023).

Tabla 3.9

Cantidad de nulos en la clase 1 y clase 2 para el grupo drug.

	Clase 1 (Con cardiopatía)		Clase 2 (Sin cardiopatía)	
	Column	**Nulos**	**Column**	**Nulos**
1	antivertigious_sum	0	antivertigious_sum	0
2	antiarrhythmics_mean	0	antiarrhythmics_mean	0
3	antiarrhythmics_count	0	antiarrhythmics_count	0
4	antiarrhythmics_slope	0	antiarrhythmics_slope	0
⋮	⋮	⋮	⋮	⋮
145	antiarrhythmics_sum	0	antiarrhythmics_sum	0
146	antiarrhythmics_mean	0	antiarrhythmics_mean	0
147	antiarrhythmics_count	0	antiarrhythmics_count	0
148	antiarrhythmics_slope	0	antiarrhythmics_slope	0

Nota. Todas las variables del grupo *drug* presentan cero valores nulos en ambas clases.

3.2.4.6. Columnas cero y *slope*

Durante el proceso de análisis de datos, fue fundamental garantizar que las variables incluidas en el conjunto de datos fueran relevantes y aportaran información significativa para los objetivos de la investigación. En este contexto, se identificaron columnas con todos sus valores en cero, lo que indicaba una falta de variabilidad o utilidad informativa.

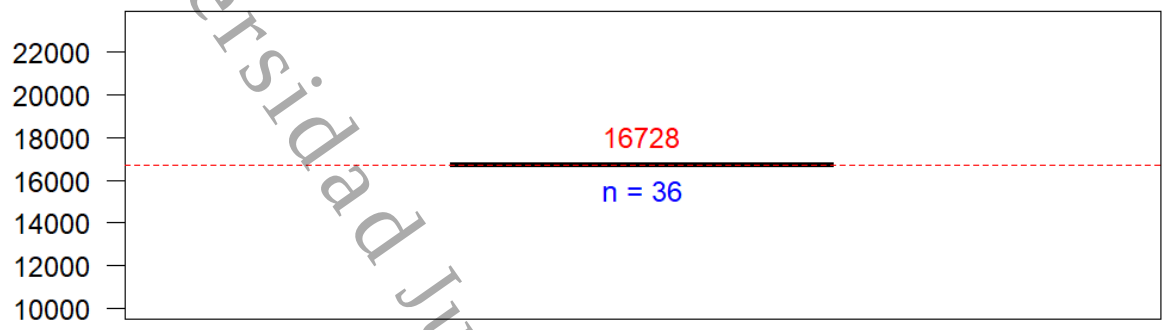
Para optimizar el análisis y enfocarnos en las variables con mayor impacto, se decidió eliminar estas columnas que no proporcionaban valor alguno. Esta acción permitió simplificar el conjunto de datos y concentrarse en las características con relevancia significativa para el estudio.

En la Figura 3.1, a través de un diagrama de caja y bigotes, se observa que el número total de columnas con valores completamente cero en la clase general fue de 36. Estas columnas contenían un total de 16,728 datos que no aportaban información útil. Por esta razón, dichas variables fueron eliminadas, quedando el conjunto de datos con 340 variables donde al menos un valor era distinto de cero y que sí ofrecen información relevante.

Es importante destacar que el 97.25 % de las columnas eliminadas pertenecían al grupo *Medical Diagnosis*, lo que equivale a 35 variables. El restante 2.75 % correspondía al grupo *Drugs*, del cual solo se eliminó una variable.

Figura 3.1

Cantidad de máximos ceros en la clase general.



Para el caso de las variables que contienen todas las observaciones con un valor constante de "1", únicamente se identificó la variable `diabetes_mellitus_type_2`. Esto se debe a que, para esta investigación, todos los sujetos en el estudio deben cumplir con el criterio de tener diabetes mellitus tipo 2, por lo que esta variable no aporta información adicional valiosa al análisis.

En los grupos *Measurements* y *Drugs*, se decidió descartar las variables relacionadas con la pendiente (*slope*). Aunque estas variables representan un valor de pendiente, no se encontró una relación significativa con otras variables dentro de los respectivos grupos. Por este motivo, estas variables no contribuyen con información relevante para el estudio y fueron eliminadas del conjunto de datos.

3.2.4.7. Variable count_cx_w con valor *null* o cero

La variable `count_cx_w` hace referencia al número de visitas realizadas por un paciente en un periodo de dos años (por ventana), con un rango que va de 0 a 80 visitas. Sin embargo, un paciente que no tiene visitas registradas en una ventana no puede aportar datos nuevos relevantes, ya que, según la documentación del *dataset*, si un paciente ya contaba con enfermedades crónicas

diagnosticadas, estos datos eran copiados automáticamente para las siguientes ventanas. Esto implica que las observaciones sin visitas reflejan datos duplicados basados en el historial médico y no información nueva.

Por este motivo, se procedió a eliminar las observaciones con visitas nulas o igual a cero, lo que resultó en la eliminación de 136 observaciones correspondientes a 51 pacientes. Cabe destacar que, aunque estas observaciones fueron descartadas, los registros de estos pacientes con visitas confirmadas y datos reales se conservaron.

En la Tabla 3.10 se presentan las CURPs anonimizadas que estaban repetidas, además de la cantidad de veces que se repiten, lo cual corresponde a la cantidad de valores nulos registrados en la variable.

Después de este proceso, el *dataset* quedó con 16,592 observaciones. No obstante, estas observaciones se reducirán aún más debido a la necesidad de balancear nuevamente el *dataset* tras la eliminación de dichas observaciones.

Finalmente, la variable `count_cx_w` también se eliminó, ya que, aunque un paciente pueda tener muchas visitas en un periodo de dos años, las métricas contenidas en el grupo *Measurements* operan bajo medidas de tendencia central y no dependen de la cantidad de visitas registradas. Esto no significa que la cantidad de visitas sea irrelevante, sino que, debido al funcionamiento de la red neuronal, al momento de procesar el valor de cada variable solo se puede ingresar un único valor por vez. Por lo tanto, la variable `count_cx_w` no aporta información adicional para el modelo.

Tabla 3.10

Cantidad de nulos en la variable count_cx_w.

	PacienteID	Repeticiones
1	00ba6c3292	4
2	0129733923	2
3	01f7a131a3	7
4	187ace512c	3
5	1a4bae476d	5
6	1e232f647f	3
7	1ff80c91f4	5
8	264ef9328c	4
9	299f6d5566	9
10	2a70aa7ff2	7
11	2c715d148a	1
12	35826ab163	1
13	3f8ebd5e76	7
14	45590bba1a	5
15	49d9fcaa65	1
16	4e4755338e	1
17	536687f9ce	1
18	5396dc0e23	1
19	564157735a	9
20	59fd1ebf16	1
21	59fe11e6f1	1
22	5b76d9e75d	2
23	5e48a9d0f2	4
24	5fbea703f0	3
25	62d996f03a	1
26	64f449677e	1
27	661c7c5a0a	2
28	6680303452	1
29	6719f59580	1
30	6ab99a334a	1
31	7b3356a5c5	2
32	7ba9f95a7e	2
33	8f0b2c5fe8	1
34	9627b77f0c	1
35	96ff63b6b6	1
36	9eac7af669	5
37	a9bc2f5d72	1
38	b357eed8cf	1
39	ba9717ece0	1
40	c68c858905	1
41	c79ef50743	1
42	ce2f92fbb2	1
43	d337250fa9	1
44	e0075d2672	4
45	e186587b0c	4
46	e6ec7ba4e7	1
47	ec2b21b28b	1
48	ec5b3164a6	3
49	ef5353ab0d	1
50	f52ca2de89	1
51	fbe68e4a26	8

Nota. Se muestran los pacientes con múltiples registros en la variable count_cx_w.

3.2.5. *Outliers*

En el análisis de datos, es fundamental realizar un estudio exhaustivo para identificar los datos atípicos, también conocidos como *outliers*. Estos son valores que se desvían significativamente del patrón general y pueden deberse a errores humanos o errores en las mediciones (Yaque, 1988).

Sin embargo, no siempre deben descartarse automáticamente, ya que podrían reflejar eventos reales bajo condiciones excepcionales. El tratamiento adecuado de los datos atípicos es crucial, ya que, desde un punto de vista estadístico, su presencia puede sesgar los resultados y afectar la precisión de los modelos utilizados en el análisis. Por ello, es necesario evaluar cuidadosamente su impacto y decidir si deben ser eliminados, ajustados o considerados como parte del fenómeno bajo estudio.

En el caso del grupo *Medical Diagnosis*, no se realiza un estudio de los datos atípicos debido a que las variables en este grupo son binarias. De igual manera, el análisis de datos atípicos se excluye para el grupo *Drug*, ya que las variables en este grupo reflejan la cantidad de medicamento que un paciente está consumiendo, lo cual no resulta pertinente para el tratamiento de datos atípicos.

Por tanto, el análisis de datos atípicos se enfoca en los grupos *Demographics*, *Measurements* y *Laboratories*. Dentro del grupo *Demographics*, de sus cinco variables, una es binaria (`cs_sex`), la cual indica el sexo del paciente. Además, incluye dos variables de tipo *string*: `id` y `cx_curp`, que corresponden a la identificación anonimizada del paciente y su CURP anonimizada, utilizada para llevar el conteo de visitas del paciente. Estas características no son aptas para el análisis de datos atípicos.

Por lo tanto, el análisis de datos atípicos se limita a un total de 205 variables, correspondientes a los grupos seleccionados que cuentan con información relevante para esta etapa del estudio.

3.2.5.1. *Demographics*

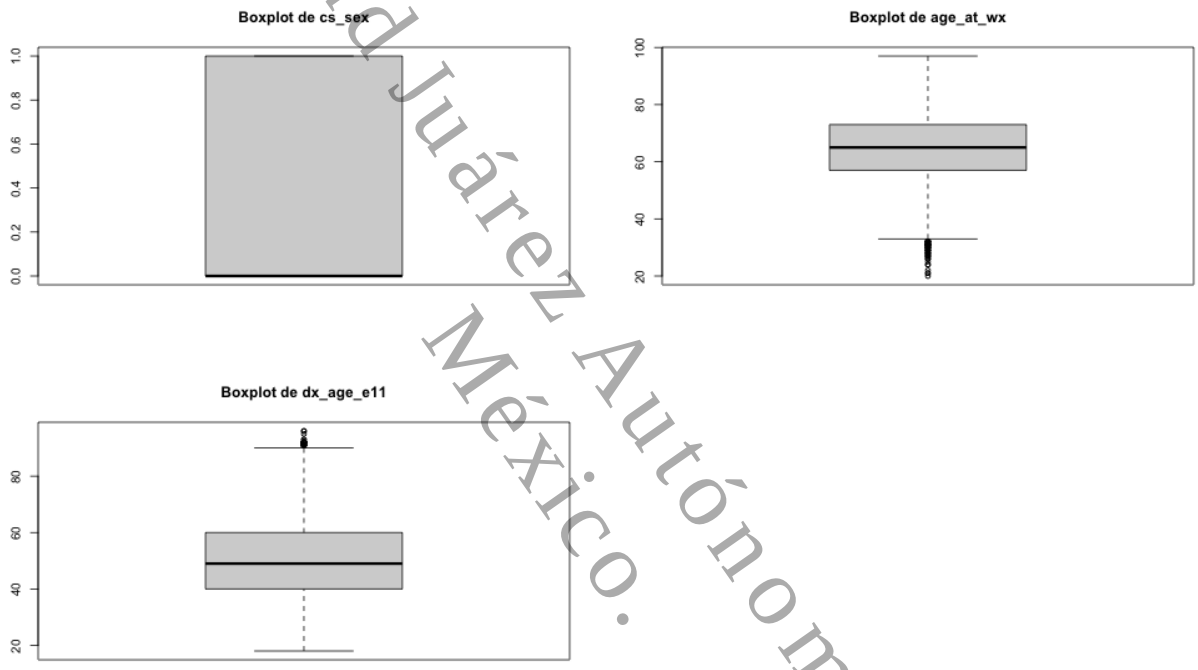
Como demostración del grupo *Demographics*, se generó un diagrama de caja y bigotes (*boxplot*) para la variable correspondiente al sexo, representada por datos binarios. Como se observa, esta variable no contiene datos atípicos, ya que sus valores están limitados a ceros y unos. Debido a

esto, no es posible identificar datos anómalos en variables binarias, lo que justifica que el análisis de datos atípicos se enfoque exclusivamente en variables numéricas continuas.

En la Figura 3.2, se puede observar que los datos numéricos continuos dentro de este grupo presentan un comportamiento distinto, permitiendo identificar la posible presencia de datos atípicos que requieren un tratamiento adicional en el análisis.

Figura 3.2

Boxplots del grupo Demographics.



Nota. Elaboración propia con base en el análisis de datos.

Si bien podría parecer que las variables `age_at_wx` y `dx_age_e11` contienen datos atípicos, es importante considerar su contexto. Estas variables hacen referencia a la "edad del paciente al final del período de tiempo para los valores `x_end`" y a la "edad al momento del diagnóstico de diabetes mellitus tipo 2", respectivamente.

Sin embargo, según la literatura, el rango de edad para desarrollar diabetes mellitus tipo 2 comienza a partir de los 20 años (Basto-Abreu et al., 2023). Por lo tanto, los valores observados dentro de estas variables pueden no representar datos anómalos, sino reflejar el rango esperado

para esta enfermedad.

3.2.5.2. *Measurements*

3.2.5.3. Variable fn_weight

La aparición de diabetes mellitus tipo 2 está asociada con la obesidad en aproximadamente el 85 % de los casos (Pérez y Vicuña, 2022). El peso corporal es una variable estrechamente relacionada con el desarrollo de esta enfermedad, y una forma general de evaluar si una persona es propensa a desarrollarla es mediante el cálculo del Índice de Masa Corporal (IMC).

En este estudio, se considera a pacientes a partir de los 20 años de edad, por lo que se utiliza la fórmula del IMC para adultos, definida como:

$$IMC = \frac{\text{Peso (kg)}}{[\text{Estatura (m)}]^2}$$

Esta fórmula toma en cuenta la altura del paciente y clasifica como obesidad un IMC superior a 30. En la Tabla 3.11 se presentan los niveles del Índice de Masa Corporal de acuerdo con la Organización Mundial de la Salud (OMS, 2007).

Tabla 3.11

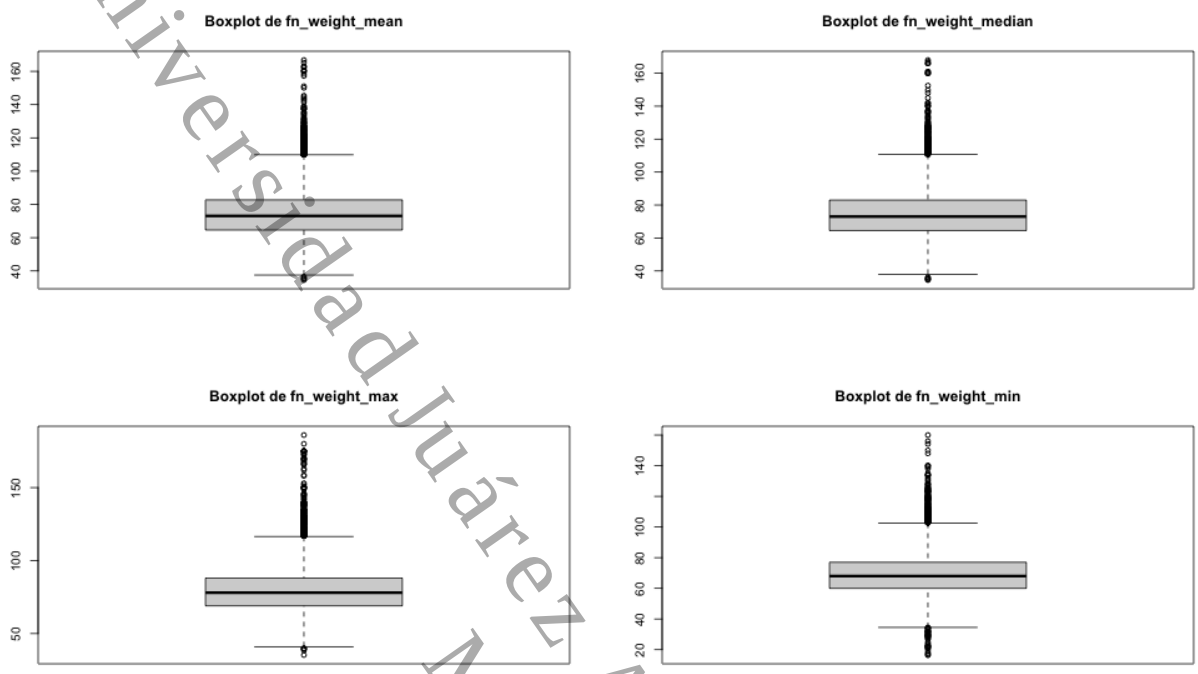
Intervalos de peso por estatura y categoría del IMC según la OMS.

	Normal		Sobrepeso		Obesidad I		Obesidad II		Obesidad III
Estatura	Min	Max	Min	Max	Min	Max	Min	Max	$\geq$
IMC	18.5	24.9	25.0	29.9	30.0	34.9	35.0	39.9	40.0
1.44	38.4	51.6	51.8	62.0	62.2	72.4	72.6	82.7	82.9
1.46	39.4	53.0	53.3	63.7	63.9	74.4	74.6	85.1	85.3
1.48	40.5	54.5	54.8	65.5	65.7	76.4	76.7	87.4	87.6
1.50	41.6	56.0	56.3	67.3	67.5	78.5	78.8	89.8	90.0
1.52	42.7	57.5	57.8	69.1	69.3	80.6	80.9	92.2	92.4
1.54	43.9	59.1	59.3	70.9	71.1	82.8	83.0	94.6	94.9
1.56	45.0	60.6	60.8	72.8	73.0	84.9	85.2	97.1	97.3
1.58	46.2	62.2	62.4	74.6	74.9	87.1	87.4	99.6	99.9
1.60	47.4	63.7	64.0	76.5	76.8	89.3	89.6	102.1	102.4
1.62	48.6	65.3	65.6	78.5	78.7	91.6	91.9	104.7	105.0
1.64	49.8	67.0	67.2	80.4	80.7	93.9	94.1	107.3	107.6
1.66	51.0	68.6	68.9	82.4	82.7	96.2	96.4	109.9	110.2
1.68	52.2	70.3	70.6	84.4	84.7	98.5	98.8	112.6	112.9
1.70	53.5	72.0	72.3	86.4	86.7	100.9	101.2	115.3	115.5
1.72	54.7	73.7	74.0	88.5	88.8	103.2	103.5	118.0	118.3
1.74	56.0	75.4	75.7	90.5	90.8	105.7	106.0	120.8	121.1
1.76	57.3	77.1	77.4	92.6	92.9	108.1	108.4	123.6	123.9
1.78	58.6	78.9	79.2	94.7	95.1	110.6	110.9	126.4	126.7
1.80	59.9	80.7	81.0	96.9	97.2	113.1	113.4	129.3	129.6
1.82	61.3	82.5	82.8	99.0	99.4	115.6	115.9	132.2	132.5
1.84	62.6	84.3	84.6	101.2	101.6	118.2	118.5	135.1	135.4

Nota. Los valores indican los rangos de peso (en kilogramos) correspondientes a cada categoría del índice de masa corporal (IMC), en función de la estatura del paciente. Clasificación según la Organización Mundial de la Salud (OMS).

Figura 3.3

Variable fn_weight.



El gráfico de cajas correspondiente a `fn_weigh_mean` muestra valores aparentemente atípicos. A continuación, se describen los casos específicos de peso mínimo y máximo que se encuentran fuera de los límites inferior y superior:

En el caso del peso mínimo, se observa a una mujer con una altura de 1.41 metros y un peso de 34.54 kilos, lo que resulta en un IMC de 17.37. Este caso corresponde a una paciente con 29 años de diagnóstico de diabetes y registrada como la número 15,665 en el *dataset*. Un IMC menor a 18.5 indica desnutrición; sin embargo, la diferencia entre su peso actual y su peso ideal, que sería de 36.78 kilos, es de apenas 2.24 kilos. Por esta razón, este peso no se considera atípico.

En cuanto al peso máximo, se identifica a un hombre con una altura de 1.71 metros y un peso de 166.79 kilos, lo que equivale a un IMC de 57.04. Este paciente tiene 3 años de diagnóstico de diabetes y está registrado como el número 3,142 en el *dataset*. Un IMC mayor a 40 clasifica como obesidad de grado 3 (obesidad mórbida). En este caso, la diferencia entre su peso actual y su peso ideal, que sería de 72.81 kilos, es de 93.98 kilos. Aunque este valor puede parecer extremo,

el paciente ha mantenido un comportamiento constante en su peso, lo que indica que estos datos reflejan una condición real y no anomalías.

Dado que los datos en ambos casos reflejan condiciones reales y consistentes con los pacientes, las variables de peso no serán tratadas como datos atípicos. El tratamiento de estas variables se enfocará en los datos nulos en una etapa posterior.

En la Tabla 3.12, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con el peso. Tomando como referencia los valores normales, se llegó a la decisión de no considerar estas variables como datos atípicos.

Tabla 3.12

Valores mínimos y máximos de las variables relacionadas con el peso.

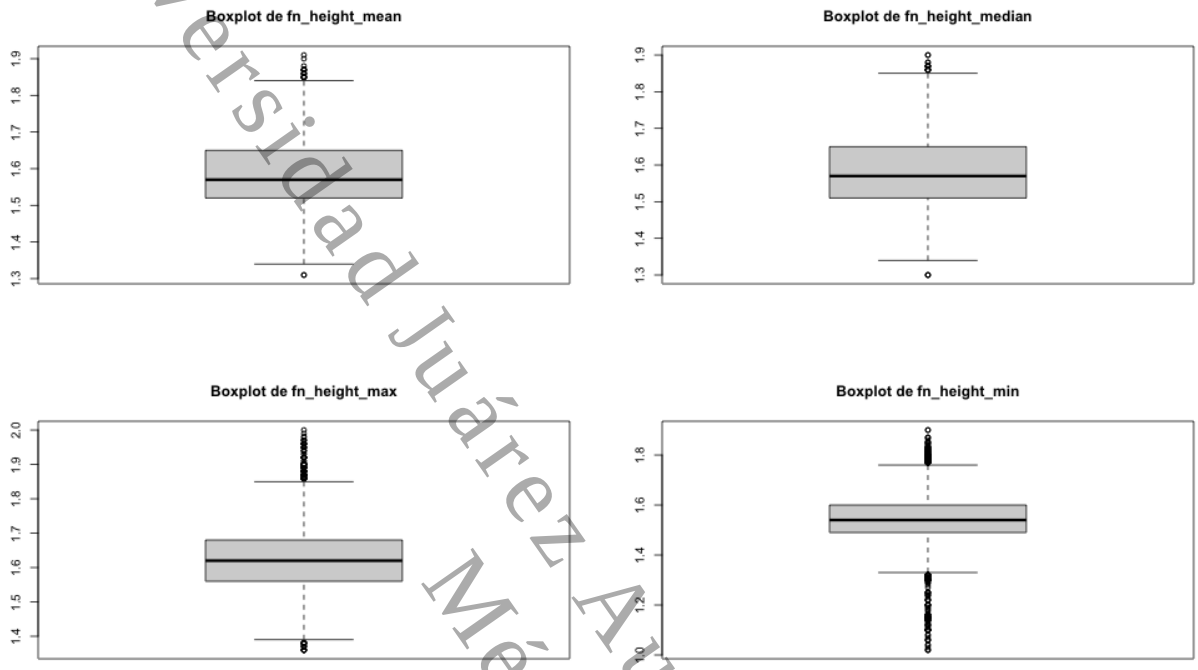
Variable	Valor mínimo	Valor máximo
fn_weight_median	34.5	168.0
fn_weight_min	16.6	160.0
fn_weight_max	35.3	186.0

Nota. Los valores están expresados en kilogramos.

3.2.5.4. Variable fn_height

Figura 3.4

Variable fn_height.



En el caso de la altura mínima, se observa a una mujer con una altura de 1.31 metros y un peso de 55.15 kilos, lo que resulta en un IMC de 32.14. Este caso corresponde a una paciente con 10 años de diagnóstico de diabetes, registrada como el número 5,148 en el *dataset*.

En cuanto a la altura máxima, se identifica a un hombre con una altura de 1.91 metros y un peso de 95.26 kilos, lo que equivale a un IMC de 26.11. Este paciente tiene 8 años de diagnóstico de diabetes y está registrado como el número 5,347 en el *dataset*.

En ambos casos, el IMC está dentro de los parámetros normales para una persona con diabetes, lo que indica que estos datos, aunque inicialmente parecen atípicos, son completamente normales y reflejan condiciones reales de los pacientes.

Las variables relacionadas con la altura son determinantes en el análisis de personas con diabetes, ya que al observarlas desde la perspectiva del IMC es posible evaluar si los datos son reales

o no. A partir de este análisis, se concluye que las variables contienen datos reales y valores normales.

En la Tabla 3.13, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la altura. Tomando como referencia los valores normales, se decidió no tratar estas variables como datos atípicos.

Tabla 3.13

Valores mínimos y máximos de las variables relacionadas con la altura.

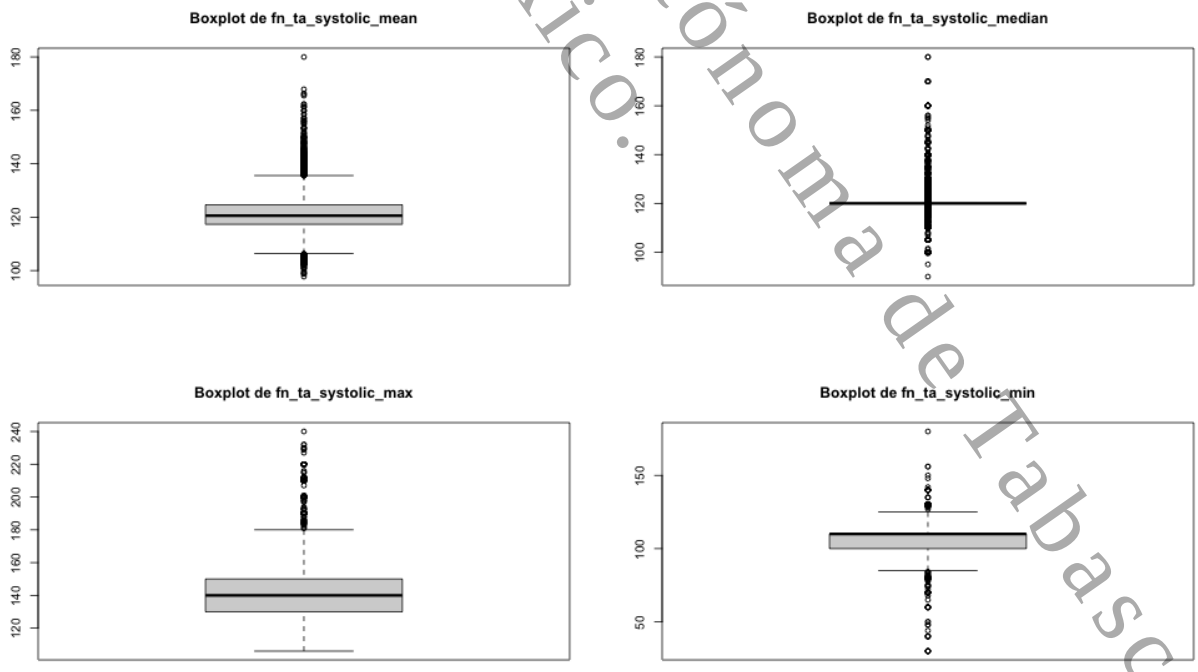
Variable	Valor mínimo	Valor máximo
fn_height_median	1.30	1.90
fn_height_min	1.02	1.90
fn_height_max	1.36	2.00

Nota. Los valores están expresados en metros.

3.2.5.5. Variable fn_ta_systolic

Figura 3.5

Variable fn_ta_systolic.



La variable `fn_ta_systolic` hace referencia al valor de la presión arterial sistólica, que se mide en milímetros de mercurio (mmHg). Este valor representa la presión en las arterias cuando el corazón bombea sangre y es el número superior en una lectura de presión arterial (AHA, 2024).

En la Tabla 3.14 se presentan los valores correspondientes a la presión arterial sistólica y diastólica, utilizados como referencia en este estudio.

Tabla 3.14

Valores para la presión sistólica y diastólica.

Categoría de la presión arterial	Sistólica (mm Hg)		Diastólica (mm Hg)
Normal	Menos de 120	y	Menos de 80
Elevada	120–129	y	Menos de 80
Presión arterial alta (hipertensión) nivel 1	130–139	o	80–89
Presión arterial alta (hipertensión) nivel 2	140 o más alta	o	90 o más alta
Crisis de hipertensión (consulte al médico de inmediato)	Más de 180	y/o	Más de 120

Nota. Adaptado de guías clínicas para el diagnóstico de hipertensión arterial.

El valor mínimo de la presión sistólica corresponde a una mujer con una altura de 1.6 metros y un peso de 102.32 kilos, lo que resulta en un IMC de 39.97. Este caso corresponde a una paciente con 18 años de diagnóstico de diabetes, registrada como el número 1,852 en el *dataset*, cuya presión sistólica es de 97.73 mmHg.

Por otro lado, el valor máximo de la presión sistólica pertenece a un hombre con una altura de 1.69 metros y un peso de 71 kilos, lo que equivale a un IMC de 24.86. Este paciente tiene 3 años de diagnóstico de diabetes y está registrado como el número 2,551, con una presión sistólica de 180 mmHg.

Aunque estos valores son señalados como *outliers*, ambos se encuentran dentro de los parámetros normales para una persona. Por esta razón, el tratamiento de los datos se enfocará en la imputación en lugar de considerarlos como datos atípicos.

En la Tabla 3.15, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la presión sistólica. Tomando como referencia los valores normales, se concluye que estas variables no serán tratadas como datos atípicos.

Tabla 3.15

Valores mínimos y máximos de las variables relacionadas con la presión sistólica.

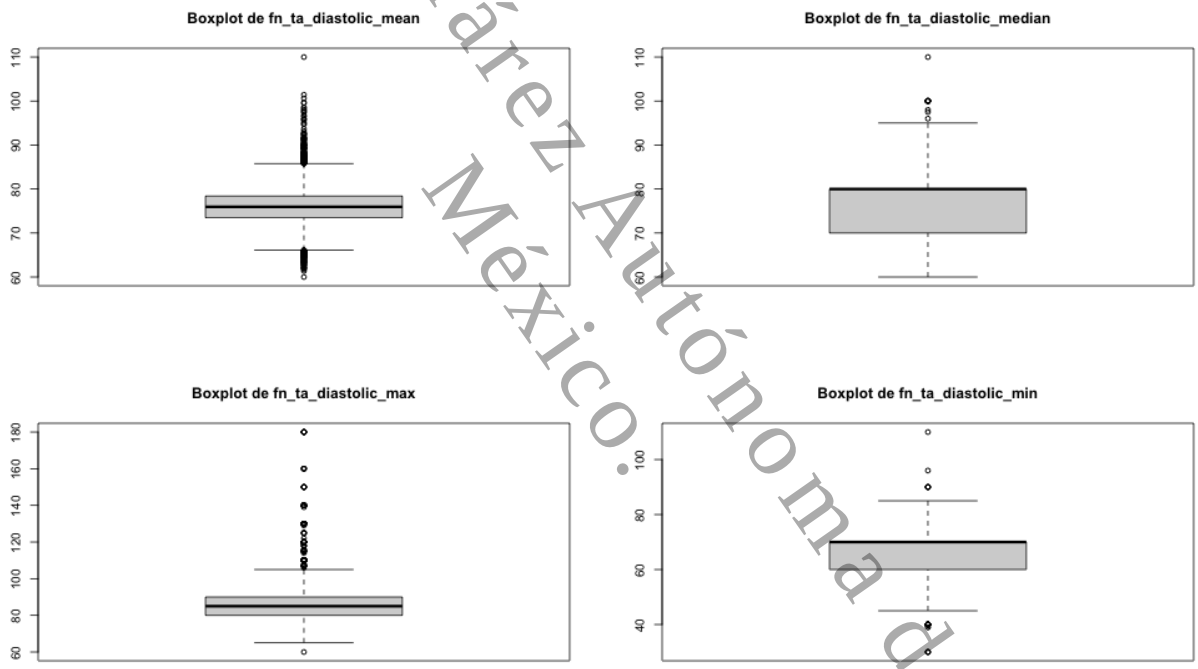
Variable	Valor mínimo	Valor máximo
fn_ta_systolic_median	90	180
fn_ta_systolic_min	30	180
fn_ta_systolic_max	106	240

Nota. Los valores están expresados en milímetros de mercurio (mm Hg).

3.2.5.6. Variable fn_ta_diastolic

Figura 3.6

Variable fn_ta_diastolic.



El valor mínimo de la presión diastólica corresponde a una mujer con una altura de 1.5 metros y un peso de 87 kilos, lo que resulta en un IMC de 38.67. Este caso pertenece a una paciente con 3 años de diagnóstico de diabetes, registrada como el número 6,644 en el *dataset*, cuya presión diastólica es de 60 mmHg.

Por otro lado, el valor máximo de la presión diastólica corresponde a un hombre con una altura de

1.69 metros y un peso de 71 kilos, lo que equivale a un IMC de 24.86. Este paciente tiene 3 años de diagnóstico de diabetes y está registrado como el número 2,551, con una presión diastólica de 110 mmHg.

Ambos valores representan datos reales que se encuentran dentro de los parámetros normales para la presión diastólica. Por lo tanto, aunque inicialmente puedan parecer valores atípicos, no serán ajustados.

En la Tabla 3.16, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la presión diastólica. Tomando como referencia los valores normales, se concluye que estas variables no serán tratadas como datos atípicos.

Tabla 3.16

Valores mínimos y máximos de las variables relacionadas con la presión diastólica.

Variable	Valor mínimo	Valor máximo
fn_ta_systolic_median	60	110
fn_ta_systolic_min	30	110
fn_ta_systolic_max	60	180

Nota. Los valores están expresados en milímetros de mercurio (mm Hg).

3.2.5.7. Variable in_heart_rate

Esta variable hace referencia al "valor de frecuencia cardíaca ([U]/min)" según el diccionario de datos. Las unidades típicas para medir la frecuencia cardíaca son latidos por minuto (bpm, por las siglas en inglés *beats per minute*).

De acuerdo con los valores normales establecidos a partir de los 20 años, y considerando el estado de reposo, los parámetros se presentan en la Tabla 3.17 (FEC, 2024b).

Tabla 3.17

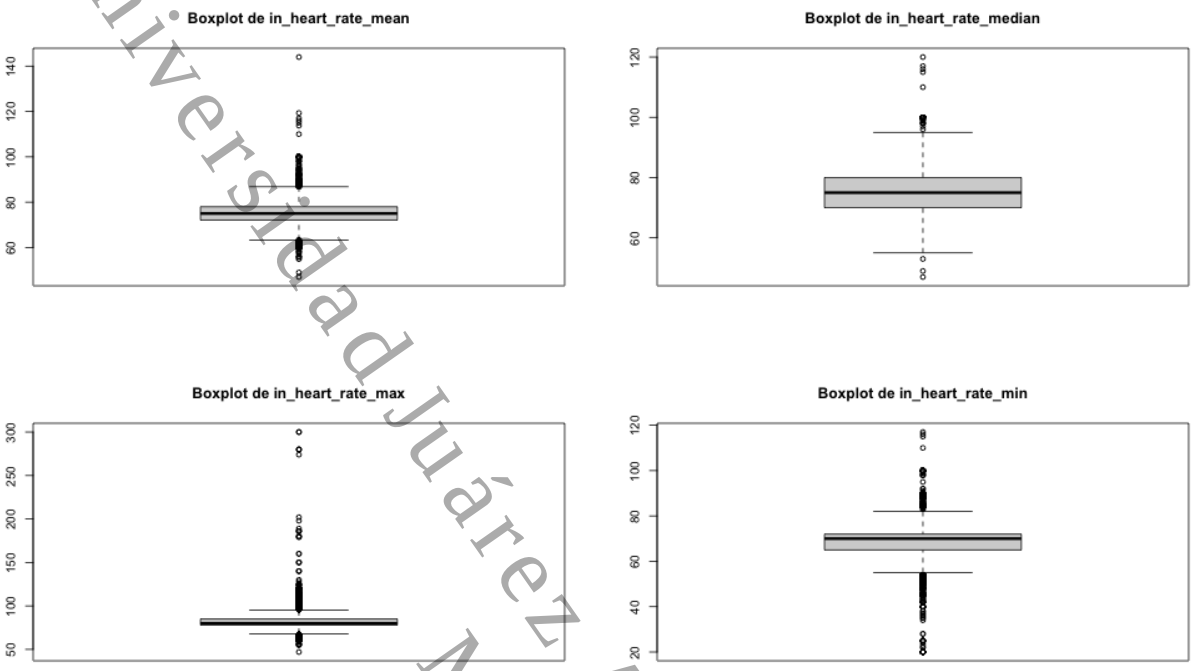
Valores de referencia para la frecuencia cardíaca.

Categoría de los latidos por minuto	Pulsaciones por minuto
Bradicardia	Menos de 60
Ordinaria	60–100
Taquicardia	Más de 100

Nota. Las categorías están basadas en los rangos estándar clínicos de frecuencia cardíaca en adultos.

Figura 3.7

Variable in_heart_rate.



El valor mínimo de frecuencia cardíaca corresponde a un hombre con una altura de 1.65 metros y un peso de 124 kilos, lo que resulta en un IMC de 45.55. Este paciente tiene 52 años, está registrado como el número 9,467 en el *dataset* y presenta una frecuencia cardíaca de 47 latidos por minuto. Además, se observó que este paciente está tomando medicamentos antihipertensivos.

El valor máximo de frecuencia cardíaca corresponde a una mujer con una altura de 1.46 metros y un peso de 78.68 kilos, lo que equivale a un IMC de 36.91. Esta paciente tiene 64 años, está registrada como el número 10,558 en el *dataset* y presenta una frecuencia cardíaca de 144 latidos por minuto.

Según la fórmula para calcular la frecuencia cardíaca máxima ($FCM = 220 -$ edad) (Miragaya y Magri, 2016), el valor máximo permitido para el paciente mencionado sería:

$$220 - 64 = 156 \text{ latidos por minuto.}$$

Por lo tanto, su frecuencia cardíaca de 144 se encuentra dentro de los parámetros normales y no se considera un valor atípico.

Por otro lado, la bradicardia puede ser causada por diversas enfermedades o medicamentos administrados a los pacientes. Los medicamentos que reducen la frecuencia cardíaca incluyen digoxina, bloqueadores beta adrenérgicos y antagonistas del calcio (Hayes, 2005). Los betabloqueantes, en particular, son un grupo amplio de medicamentos utilizados para tratar condiciones como angina de pecho, hipertensión arterial, prevención de infartos, arritmias, insuficiencia cardíaca, miocardiopatía hipertrófica, glaucoma y ansiedad (Rodríguez y Mármol, 2017).

Dado que el paciente con el valor mínimo de frecuencia cardíaca está tomando medicamentos antihipertensivos, su valor de 47 latidos por minuto también se considera dentro de lo esperado y no es un valor atípico.

En la Tabla 3.18, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la frecuencia cardíaca. Tomando como referencia los valores normales, se concluye que estas variables no serán tratadas como datos atípicos.

Tabla 3.18

Valores mínimos y máximos registrados para las variables relacionadas con la frecuencia cardíaca.

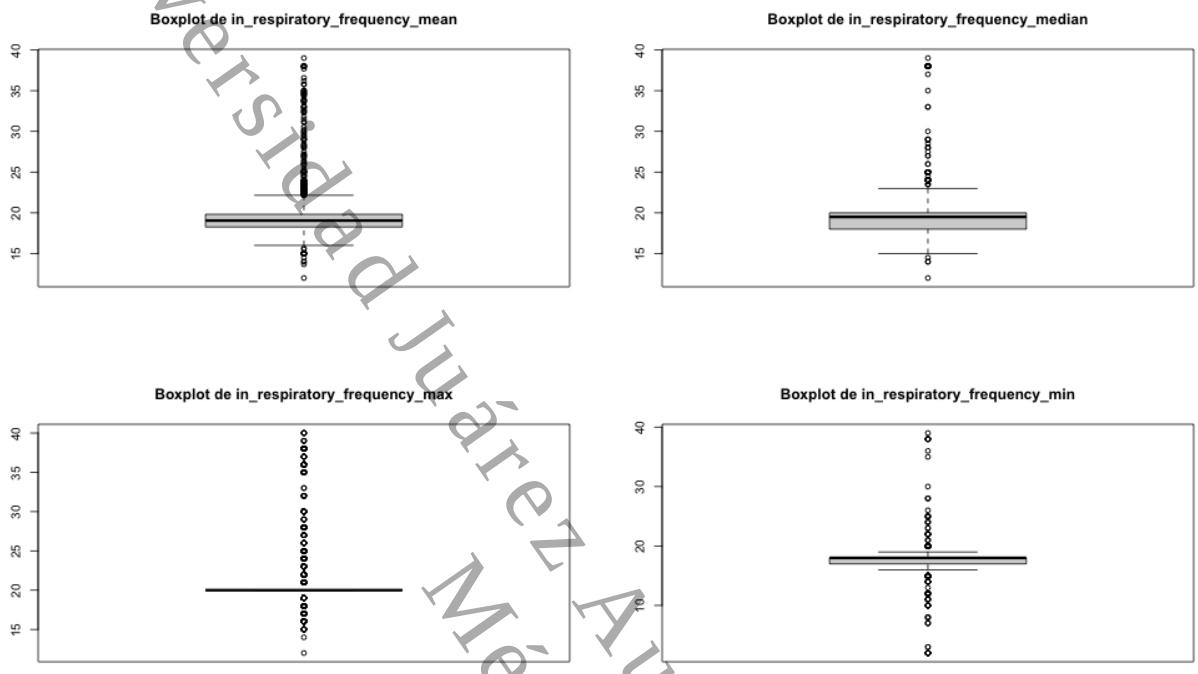
Variable	Valor mínimo	Valor máximo
in_heart_rate_median	47	120
in_heart_rate_min	20	117
in_heart_rate_max	47	300

Nota. Se presentan los valores extremos de variables que describen la frecuencia cardíaca por paciente en diferentes momentos.

3.2.5.8. Variable in_respiratory_frequency

Figura 3.8

Variable in_respiratory_frequency.



El valor mínimo de frecuencia respiratoria corresponde a un hombre con una altura de 1.78 metros y un peso de 157.22 kilos, lo que resulta en un IMC de 45.55. Este paciente tiene 53 años y está registrado como el número 11,873 en el *dataset*, con una frecuencia respiratoria de 12 respiraciones por minuto.

Por otro lado, el valor máximo de frecuencia respiratoria corresponde a una mujer con una altura de 1.53 metros y un peso de 72.84 kilos, lo que equivale a un IMC de 31.12. Esta paciente tiene 68 años, está registrada como el número 1,208 y presenta una frecuencia respiratoria de 39 respiraciones por minuto. Además, se identificó que esta paciente presenta “enfermedades del sistema respiratorio”, lo que explica el valor elevado de su frecuencia respiratoria.

De acuerdo con los valores establecidos como normales para la frecuencia respiratoria en estado de reposo, que oscilan entre 12 y 20 respiraciones por minuto (Rowden, 2022), el valor mínimo observado se encuentra dentro de los parámetros esperados. Por su parte, el valor máximo puede

ser atribuido a la condición respiratoria del paciente, lo que también lo excluye como un dato atípico.

En la Tabla 3.19, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la frecuencia respiratoria. Tomando como referencia los valores normales, se concluye que estas variables no serán tratadas como datos atípicos.

Tabla 3.19

Valores mínimos y máximos registrados para las variables relacionadas con la frecuencia respiratoria.

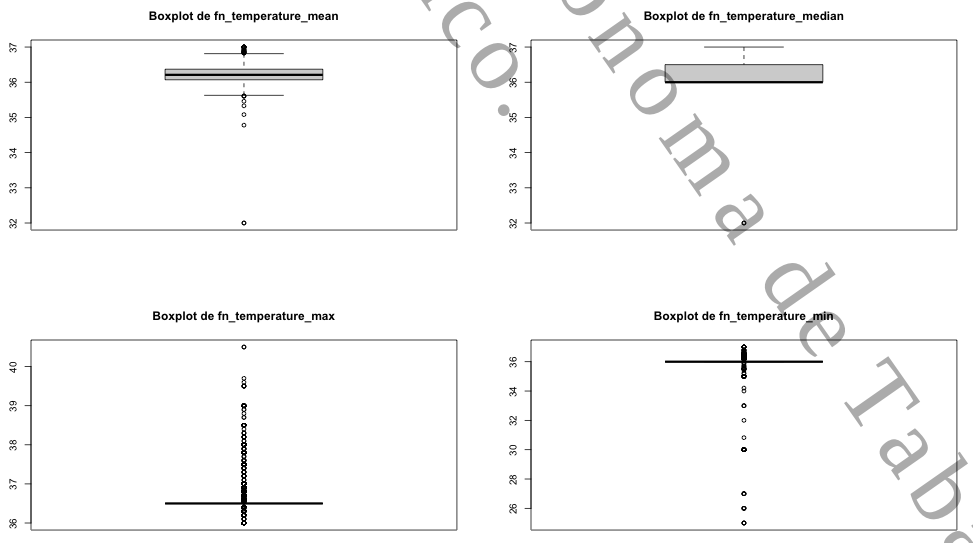
Variable	Valor mínimo	Valor máximo
in_respiratory_frequency_median	12	39
in_respiratory_frequency_min	2	39
in_respiratory_frequency_max	12	40

Nota. Se muestran los valores extremos registrados para las variables de frecuencia respiratoria por paciente.

3.2.5.9. Variable fn_temperature

Figura 3.9

Variable fn_temperature.



Los rangos de temperatura se muestran en tabla 3.20 (Ayala, 2007).

Tabla 3.20

Categorías clínicas de la temperatura corporal en grados centígrados.

Categoría de la temperatura	Rango (℃)
Hipotermia	28–35
Normal	35.8–37.2
Fiebre	38–42

Nota. Clasificación basada en valores clínicos comunes de temperatura corporal.

La hipotermia puede ser causada por diversas enfermedades y afecciones (Blondin, 2014). Estas incluyen exposición ambiental, *shock*, infecciones, trastornos metabólicos como hipotiroidismo, insuficiencia suprarrenal y encefalopatía de Wernicke, desnutrición, así como toxicidad por alcohol o drogas.

El valor mínimo de temperatura corporal corresponde a un hombre con una altura de 1.67 metros y un peso de 89.12 kilos, lo que resulta en un IMC de 31.96. Este paciente tiene 81 años y está registrado como el número 2,631 en el *dataset*. Su temperatura corporal es de 32℃ y presenta registro positivo para “enfermedades endocrinas, nutricionales y metabólicas”.

Por otro lado, el valor máximo de temperatura corresponde a una mujer con una altura de 1.58 metros y un peso de 73.02 kilos, lo que equivale a un IMC de 29.25. Esta paciente tiene 53 años y está registrada como el número 2,029, con una temperatura corporal de 37℃.

Ambos valores, aunque señalados como *outliers*, se consideran normales. El valor mínimo refleja un caso de hipotermia relacionado con las condiciones médicas del paciente, mientras que el valor máximo representa un caso de fiebre leve dentro de los parámetros esperados.

En la Tabla 3.21, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la temperatura corporal. Tomando como referencia los valores normales, se concluye que estas variables no serán tratadas como datos atípicos.

Tabla 3.21

Valores máximos y mínimos de las variables relacionadas con la temperatura corporal.

Variable	Valor mínimo	Valor máximo
fn_temperature_median	32	37
fn_temperature_min	25	37
fn_temperature_max	36	108

Nota. Las variables se refieren a medidas agregadas de temperatura corporal obtenidas durante el seguimiento clínico.

3.2.5.10. *Laboratories*

3.2.5.11. Variable in_glucose

Esta variable hace referencia al "valor medio de glucosa en sangre (mg/dL)", un indicador crucial en el manejo de la diabetes mellitus tipo 2. Los valores normales para las personas con diabetes tipo 2 se presentan en la Tabla 3.22 (Almanza et al., 2017; CUN, 2024; B. Huang et al., 2022; MedlinePlus, 2024; Piątkiewicz, 2016).

Tabla 3.22

Valores de la glucosa en pacientes con diabetes mellitus tipo 2 (DMT2)

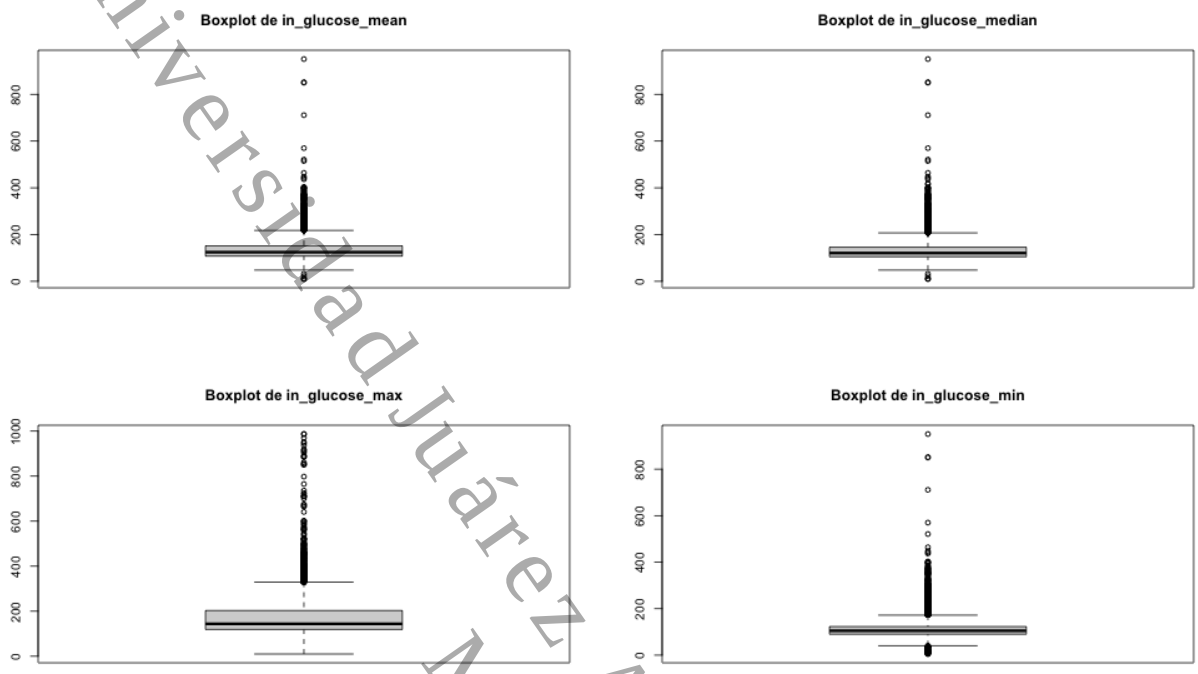
Categoría de glucosa en sangre	Glucosa (mg/dL)
Hipoglucemia nivel 2	20 – 54.05
Hipoglucemia nivel 1	54.05 – 70.26
Normal	80 – 130
Hiperglucemia	180 – 600
Hiperglucemia severa	Más de 600

Nota. Clasificación basada en rangos comunes de glucosa para personas con DMT2, útil para evaluación clínica.

La hipoglucemia es una de las complicaciones agudas más frecuentes y temidas en el tratamiento de la diabetes mellitus tipo 2 (Almanza et al., 2017). Además, se ha documentado un caso extremo de un paciente con un nivel de glucosa plasmática de 1,897 mg/dL al momento de su ingreso a la sala de emergencias (Sato et al., 2020).

Figura 3.10

Variable in_glucose.



El valor mínimo de glucosa en sangre corresponde a una mujer con una altura de 1.47 metros y un peso de 47.15 kilos, lo que resulta en un IMC de 21.82. Esta paciente tiene 88 años y está registrada como el número 4,618 en el *dataset*, con un nivel de glucosa en sangre de 10 mg/dL.

Por otro lado, el valor máximo de glucosa en sangre corresponde a una mujer con una altura de 1.56 metros y un peso de 67.96 kilos, lo que equivale a un IMC de 27.93. Esta paciente tiene 70 años y está registrada como el número 12,565, con un nivel de glucosa en sangre de 951 mg/dL.

Aunque los diagramas de caja y bigotes sugieren que existen datos atípicos en las mediciones de glucosa en sangre, se concluyó que estos valores son reales. El valor mínimo, cercano a 20 mg/dL, representa un rango donde los pacientes pueden entrar en coma hipoglucémico, mientras que el valor máximo refleja una hiperglucemia extrema.

En la Tabla 3.23, se presentan los valores máximos y mínimos de las otras tres variables relacionadas con la glucosa en sangre. Tomando como referencia los valores normales, se decidió no tratar estas variables como datos atípicos.

Tabla 3.23

Valores máximos y mínimos de las variables relacionadas con la glucosa en sangre de pacientes con DMT2.

Variable	Valor mínimo	Valor máximo
in_glucose_median	10	951
in_glucose_min	5	951
in_glucose_max	10	987

Nota. Valores obtenidos a partir de los registros clínicos de pacientes diagnosticados con diabetes mellitus tipo 2 (DMT2).

3.3. Preparación de los datos

3.3.1. Imputación de datos

3.3.1.1. Imputación de valores nulos como cero

Para cada una de las variables del grupo *Measurements*, existe una variable asociada que representa la desviación estándar. Esta es calculada en función del número de mediciones realizadas a cada paciente y los valores obtenidos en dichas mediciones.

Es importante señalar que, cuando solo se dispone de un único valor, la desviación estándar debe ser igual a cero, ya que no existe variabilidad en los datos con respecto a la media. Sin embargo, en algunos casos, estos valores aparecen representados como nulos, lo cual no es correcto. Esto se puede observar en el ejemplo presentado en la Figura 3.11, que utiliza los datos de la variable `fn_ta_systolic` como referencia.

Figura 3.11

Variable fn_ta_systolic.

fn_ta_systolic_max	fn_ta_systolic_min	fn_ta_systolic_count	fn_ta_systolic_std
130	130	1	*NA*
130	130	1	*NA*

Nota. Elaboración propia con base en el análisis de datos.

La tabla 3.24 muestra la cantidad de valores nulos encontrados en cada una de las variables de desviación estándar del grupo *Measurements* antes de la imputación a cero.

Tabla 3.24

Cantidad de nulos en la variable de desviación estándar del grupo Measurements.

Variable	Cantidad de nulos (donde Variable_count = 1 y Variable_std = NA)
fn_weight	31
fn_height	31
fn_temperature	266
fn_ta_systolic	40
fn_ta_diastolic	40
in_heart_rate	1847
in_respiratory_frequency	1894

Nota. Se consideran nulos aquellos valores cuya desviación estándar es NA y cuya cantidad de registros es igual a 1.

3.3.1.2. Imputación de datos nulos

En el ámbito de la investigación, la calidad de los datos es un aspecto fundamental para garantizar la validez y fiabilidad de los resultados obtenidos. En la Tabla 3.25, se presentan los nombres de las variables y la cantidad de datos nulos que fueron imputados. Estas variables pertenecen a los grupos *Measurements* y *Laboratories*.

Cabe destacar que la mayor cantidad de datos nulos, que asciende a 3,551, equivale al 21.4 % de valores faltantes, lo que subraya la importancia de este proceso para preservar la integridad del análisis.

Tabla 3.25

Cantidad de valores nulos imputados por variable.

Variable	Nulos imputados
fn_weight_mean	7
fn_weight_median	7
fn_weight_max	7
fn_weight_min	7
fn_weight_count	7
fn_weight_std	7
fn_height_mean	7
fn_height_median	7
fn_height_max	7
fn_height_min	7
fn_height_count	7
fn_height_std	7
fn_ta_systolic_mean	18
fn_ta_systolic_median	18
fn_ta_systolic_max	18
fn_ta_systolic_min	18
fn_ta_systolic_count	7
fn_ta_systolic_std	18
fn_ta_diastolic_mean	18
fn_ta_diastolic_median	18
fn_ta_diastolic_max	18
fn_ta_diastolic_min	18
fn_ta_diastolic_count	7
fn_ta_diastolic_std	18
in_heart_rate_mean	479
in_heart_rate_median	479
in_heart_rate_max	479
in_heart_rate_min	479
in_heart_rate_count	7
in_heart_rate_std	479
in_respiratory_frequency_mean	506
in_respiratory_frequency_median	506
in_respiratory_frequency_max	506
in_respiratory_frequency_min	506
in_respiratory_frequency_count	7
in_respiratory_frequency_std	506
fn_temperature_mean	53
fn_temperature_median	53
fn_temperature_max	53
fn_temperature_min	53
fn_temperature_count	7
fn_temperature_std	53
in_glucose_mean	3551
in_glucose_median	3551
in_glucose_max	3551
in_glucose_min	3551
in_glucose_count	7

Para la imputación de los datos, se utilizó una técnica multivariada iterativa, que es especialmente eficaz para gestionar valores faltantes en conjuntos de datos complejos (qu4nt y scikit-learn, 2021). Este enfoque fue implementado mediante cinco métodos distintos: regresión lineal, bosques aleatorios, árboles de decisión, vecinos más cercanos y máquinas de soporte vectorial.

Cada uno de estos métodos aporta perspectivas y fortalezas únicas al proceso de imputación, lo que asegura una mayor precisión y consistencia en los datos imputados. La combinación de múltiples enfoques no solo minimiza los posibles sesgos asociados con el uso de un único método, sino que también mejora la robustez de los resultados obtenidos.

3.3.1.3. Balanceo de clases

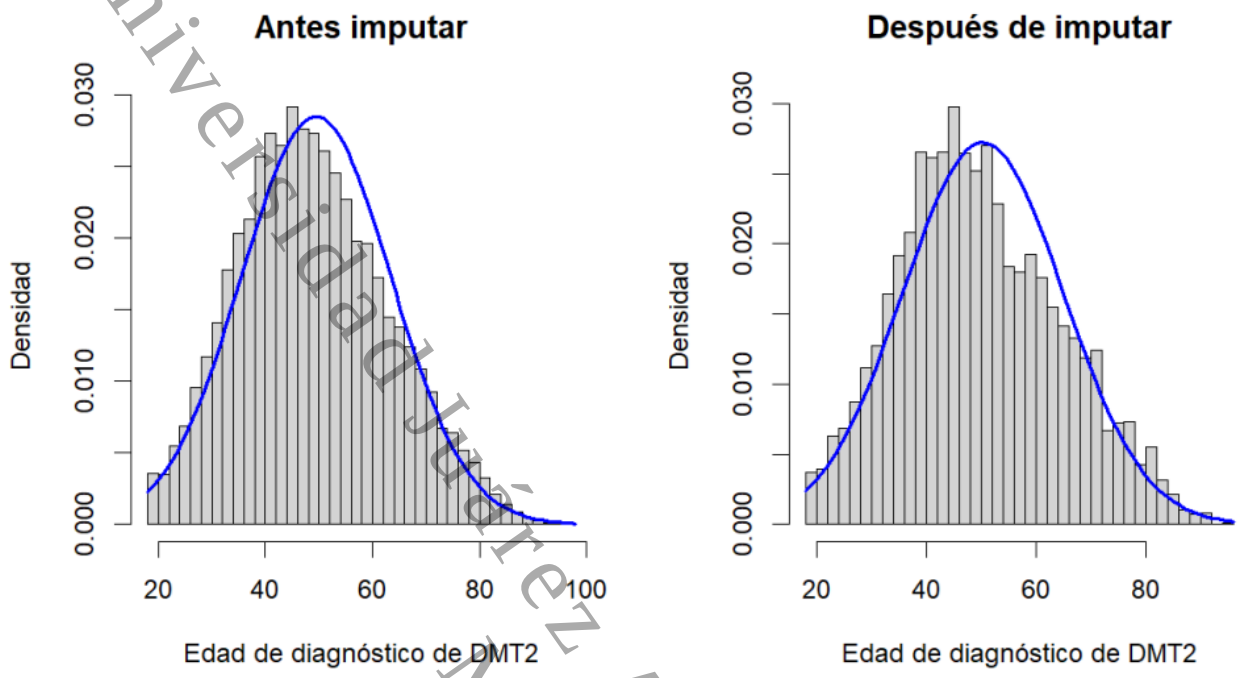
Una vez finalizada la imputación de datos, se procedió a realizar un proceso de balanceo de clases para garantizar una representación equitativa de las mismas. Este paso es fundamental para evitar sesgos en los modelos de clasificación y asegurar que cada clase esté adecuadamente representada en el análisis.

Como resultado de este proceso, se obtuvo un conjunto de datos equilibrado con 8,228 observaciones por clase (personas diabéticas con cardiopatía isquémica crónica y personas diabéticas sin cardiopatía isquémica crónica), alcanzando un total de 16,456 observaciones listas para el análisis posterior. Este balanceo asegura que los datos estén preparados de manera óptima para las siguientes etapas del estudio, facilitando la obtención de resultados más precisos y confiables.

Además, se observó que los datos conservaron el comportamiento normal esperado para una distribución de probabilidad. Esto se puede apreciar en la Figura 3.12, que muestra la distribución de la variable `dx_age_e11`, correspondiente a la edad en la que se diagnosticó la diabetes mellitus tipo 2.

Figura 3.12

Comportamiento Normal; Edad de Diagnóstico de Diabetes Mellitus Tipo 2.



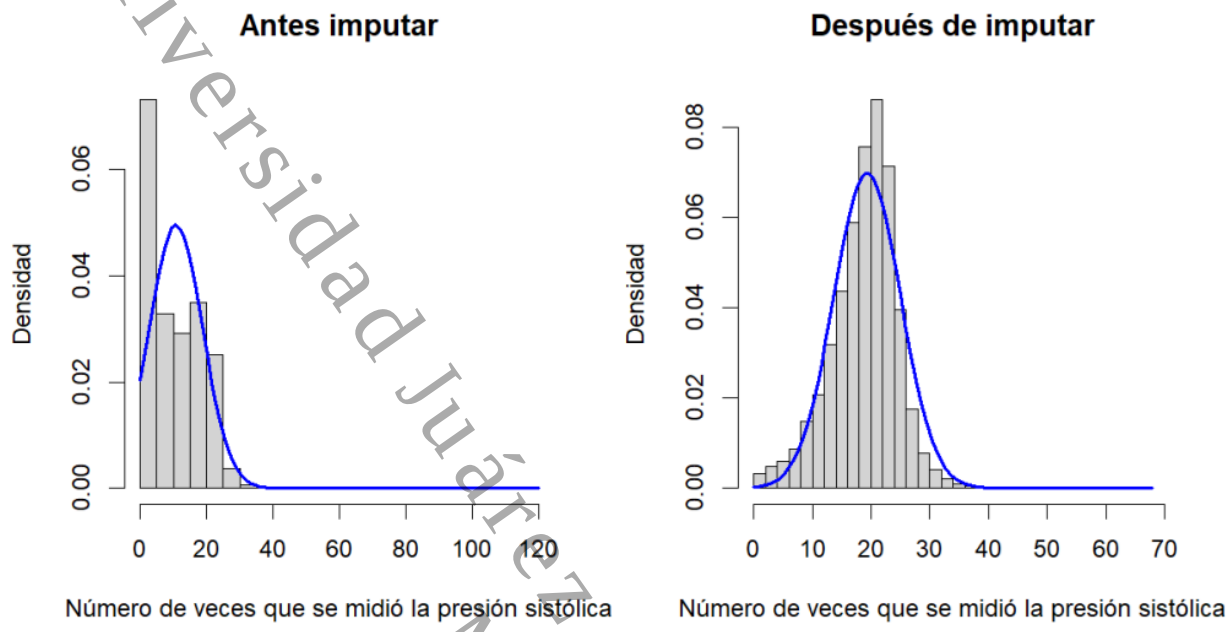
Nota. Elaboración propia con base en el análisis de datos.

Finalmente, se llevó a cabo un análisis comparativo del comportamiento de las variables antes y después del proceso de imputación. En particular, la Figura 3.13 muestra el caso de la variable `fn_ta_systolic_count`. Inicialmente, esta seguía un comportamiento tipo gamma; sin embargo, tras la imputación y el balanceo, presenta una tendencia hacia una distribución normalizada.

Este cambio refleja el impacto positivo de las técnicas empleadas, mejorando la distribución de los datos y asegurando una mayor calidad en el análisis posterior.

Figura 3.13

Comportamiento Normalizado: Número de Veces que se Midió la Presión Sistólica en Pacientes con Diabetes Mellitus Tipo 2.



3.3.1.4. Selección de características

Para identificar las variables más relevantes en el diagnóstico de la cardiopatía isquémica crónica, se implementaron cuatro métodos distintos de selección de características: *ranking* de correlación, prueba de chi-cuadrada, selección de características basada en correlación (CFS, por las siglas en inglés de *Correlation-based Feature Selection*) y correlación de Pearson. Estos métodos fueron seleccionados por su capacidad para evaluar la importancia de las variables desde diferentes enfoques estadísticos, asegurando un análisis riguroso.

Cada uno de estos enfoques proporciona una perspectiva particular sobre la relación entre las variables y el diagnóstico, lo que permite una identificación precisa y fundamentada de las características clave. Los resultados obtenidos mediante estos métodos se presentan en la Tabla 3.26.

Tabla 3.26

Variables relevantes; Dataset de Regresión Lineal.

Característica	Rank Cor	Chi-cuadrada	CFS	Correlación de Pearson
1	✓	✓	✓	✓
2	✓	✓	✓	✓
3	✓	✓	✓	✓
4	✓	✓		✓
5	✓			✓
6	✓	✓		✓
7	✓			✓
8	✓	✓		✓
9	✓	✓		✓
10	✓	✓		
11	✓	✓		✓
12	✓			✓
13	✓			✓
14	✓			✓
15	✓			✓
16	✓		✓	✓
17	✓			
18	✓			✓
19	✓	✓		
20	✓	✓		

Nota. Se muestran las características seleccionadas por cada uno de los métodos empleados.

Como se mencionó en el apartado de “imputación de datos nulos”, se utilizaron cinco métodos distintos para la imputación, generando cinco datasets diferentes. Para determinar el mejor conjunto de datos, se implementaron cuatro métodos de selección de características con el objetivo de establecer un *ranking* de atributos en cada uno de los cinco *datasets*.

Tras evaluar las métricas obtenidas, se determinó que todos los *datasets* eran aptos para esta investigación. Además, los resultados de los métodos de *ranking* coincidieron en identificar las mismas variables relevantes en cada caso. Por lo tanto, se decidió utilizar el *dataset* imputado mediante el modelo de regresión lineal.

Adicionalmente, se utilizó el método *stepwise regression* (*step*) para identificar las variables más relevantes, empleando el Criterio de Información de Akaike (*AIC*) como referencia. Este método

permitió refinar aún más la selección de características, asegurando que las variables finales fueran las más significativas, especialmente en los casos donde el *AIC* permanecía constante, garantizando así un modelo más robusto y preciso.

Las variables relevantes identificadas se presentan en el listado inferior. Asimismo, se agregó la variable dependiente "cardiopatía isquémica crónica" (`chronic_ischemic_heart_disease`) a estas 18 variables, obteniendo un *dataset* final compuesto por 19 variables.

3.3.2. Variables relevantes

1. `complications_and_ill_defined_descriptions_of_heart_disease`

   Definición: Complicaciones y descripciones mal definidas de enfermedades del corazón (CIE I51).

2. `in_heart_rate_count`

   Definición: Número de valores registrados de frecuencia cardíaca ([U]/min).

3. `dx_age_e11`

   Definición: Edad al momento del diagnóstico de diabetes mellitus tipo 2.

4. `in_respiratory_frequency_count`

   Definición: Número de valores registrados de frecuencia respiratoria ([U]/min).

5. `age_at_wx`

   Definición: Edad del paciente al final del período de tiempo para los valores (x_end).

6. `hypertensive_heart_disease`

   Definición: Enfermedad del corazón por hipertensión (CIE I11).

7. `antidiabetics_count`

   Definición: Número de medicamentos antidiabéticos registrados.

8. `antiplatelets_mean`

   Definición: Promedio de la dosis de medicamentos antiplaquetarios prescritos.

9. `antiplatelets_sum`

   Definición: Suma de la dosis de medicamentos antiplaquetarios prescritos.

10. `antiplatelets_count`

    Definición: Número de medicamentos antiplaquetarios registrados.

11. `heart_failure`

    Definición: Insuficiencia cardíaca (CIE I50).

12. `circulatory_system_diseases`

    Definición: Enfermedades del sistema circulatorio (CIE I00).

13. `in_glucose_count`

    Definición: Número de valores registrados de glucosa en sangre (mg/dL).

14. `lipid_lowering_count`

    Definición: Número de medicamentos hipolipemiantes registrados.

15. `in_heart_rate_max`

    Definición: Valor máximo de frecuencia cardíaca ([U]/min).

16. `lipid_lowering_mean`

    Definición: Promedio de la dosis de medicamentos hipolipemiantes prescritos.

17. `antidiabetics_mean`

    Definición: Promedio de la dosis de medicamentos antidiabéticos prescritos

18. `in_heart_rate_std`

    Definición: Desviación estándar de los valores de frecuencia cardíaca ([U]/min).

19. `chronic_ischemic_heart_disease` (Variable dependiente)

    Definición: Cardiopatía isquémica crónica (CIE I25).

Capítulo 4

Experimentos y resultados

En esta investigación se utilizó un TabTransformer, que es un modelo Transformer para trabajar con datos tabulares. Este modelo conserva los bloques Transformer y el mecanismo de atención, pero está diseñado para integrar características categóricas y numéricas de manera eficiente.

La elección del TabTransformer para este estudio se fundamenta en su capacidad para manejar eficientemente datos tabulares mixtos (categóricos y numéricos), característica esencial para el procesamiento de registros médicos.

Su arquitectura, que combina los beneficios del mecanismo de atención con el procesamiento específico para datos tabulares, lo hace particularmente adecuado para identificar patrones complejos en los indicadores médicos relacionados con la cardiopatía isquémica crónica. Además, su capacidad para manejar relaciones no lineales entre variables y su robustez ante datos faltantes lo convierten en una opción superior a los enfoques tradicionales de aprendizaje automático para este tipo de aplicaciones médicas.

Además del modelo TabTransformer, se seleccionaron tres algoritmos básicos de aprendizaje automático reconocidos por su amplio uso en la literatura: regresión logística, K-vecinos más cercanos (KNN, por las siglas en inglés de *k-Nearest Neighbors*) y árboles de decisión (Acharya y Shaileshbhai, 2024; Coursera, 2025). Estos métodos se eligieron por su simplicidad, eficiencia computacional y aplicabilidad demostrada a tareas de clasificación. Su inclusión permite establecer una línea base confiable y evaluar objetivamente las mejoras que ofrece una arquitectura

avanzada como TabTransformer.

4.1. Implementación del modelo

4.1.1. Arquitectura del sistema

El sistema se implementó mediante dos *scripts* principales:

- ***Script* de entrenamiento:**
  - Búsqueda de hiperparámetros mediante algoritmos genéticos.
  - Entrenamiento del modelo con los mejores hiperparámetros encontrados.
  - Validación cruzada durante el entrenamiento.
  - Guardado del modelo y métricas.
- ***Script* de validación:**
  - Evaluación del modelo con el conjunto de datos de validación.
  - Generación de análisis SHAP para interpretabilidad.
  - Cálculo de métricas finales y curva ROC.

4.1.2. Búsqueda de hiperparámetros

Se implementó un algoritmo genético para la optimización de hiperparámetros con las siguientes características:

- Población inicial de 10 individuos, seleccionada para balancear diversidad y eficiencia computacional.
- 5 generaciones de evolución, suficientes para convergencia dado el tamaño del espacio de búsqueda.
- Operadores de *crossover* y mutación adaptativa para mantener la diversidad de la población.

- Validación cruzada de 5 *folds* para cada individuo, asegurando robustez en la evaluación.

Algoritmo 1 Búsqueda de hiperparámetros mediante algoritmo genético

Input: Conjunto de datos de entrenamiento, parámetros del algoritmo.
Output: Mejor conjunto de hiperparámetros encontrado.

```
Generar población inicial aleatoria de 10 individuos
mejor individuo ← None
mejor fitness ← 0
for generacion = 1 to 5 do
    for cada individuo en población do
        scores ← [ ]
        for fold = 1 to 5 do
            modelo ← crear modelo(individuo)
            score ← entrenar y evaluar(modelo, fold)
            scores.append(score)
        end for
        fitness ← promedio(scores)
        if fitness >mejor fitness then
            mejor fitness ← fitness
            mejor individuo ← individuo
        end if
    end for
    nueva población ← [ ]
    while len(nueva población) <tamaño población do
        padres ← selección torneo(población, fitness scores)
        hijo ← crossover(padres)
        hijo ← mutar(hijo)
        nueva población.append(hijo)
    end while
    población ← nueva población
end for
return  mejor individuo
```

4.1.3. Configuración de rangos de búsqueda

Para la búsqueda de hiperparámetros, se establecieron rangos específicos basados en consideraciones teóricas y prácticas:

- ***Embedding dimension* (32-512):** Este rango se estableció considerando que:
    - Valores menores a 32 limitarían la capacidad del modelo para capturar patrones complejos en los datos médicos.
    - Valores superiores a 512 incrementarían el costo computacional sin garantizar mejoras

significativas.

    - El valor debe ser divisible por el número de *heads* del Transformer para mantener la estabilidad del modelo.

- **Transformer *layers* (1-4):** La limitación a 4 capas se fundamenta en:
    - Estudios previos que muestran que para datos tabulares médicos, más capas pueden llevar a sobreajuste.
    - La necesidad de mantener la interpretabilidad del modelo para aplicaciones médicas.
    - El balance entre capacidad de modelado y eficiencia computacional.
- **Transformer *heads* (2-16):** Los límites se establecieron para:
    - Asegurar la capacidad de capturar múltiples tipos de relaciones en los datos médicos (mínimo 2 *heads*).
    - Evitar la fragmentación excesiva de la atención (máximo 16 *heads*).
    - Mantener la eficiencia computacional del modelo.
- ***Learning rate* (1e-4 - 1e-2):** Este rango se definió para:
    - Permitir una exploración adecuada del espacio de optimización.
    - Evitar la convergencia prematura (límite inferior).
    - Prevenir la inestabilidad en el entrenamiento (límite superior).
- ***Dropout rate* (0.1-0.5):** Los límites se establecieron considerando:
    - La necesidad de prevenir el sobreajuste sin comprometer el aprendizaje.
    - Valores menores a 0.1 serían insuficientes para la regularización.
    - Valores mayores a 0.5 podrían resultar en pérdida excesiva de información.
- ***Batch size* (32-256):** Este rango se estableció considerando que:
    - Tamaños menores a 32 podrían resultar en actualizaciones ruidosas del gradiente.

    - Tamaños mayores a 256 podrían afectar la capacidad de generalización.
    - El rango permite un balance entre velocidad de entrenamiento y uso de memoria.
- **MLP *layers* (1-4):** La configuración de capas MLP se definió considerando:
    - Una capa mínima para garantizar capacidad de transformación no lineal.
    - Máximo 4 capas para evitar complejidad excesiva y *vanishing gradients*.
    - Balance entre capacidad de representación y facilidad de entrenamiento.
- ***Weight decay* (1e-5 - 1e-3):** Los límites se establecieron para:
    - Proporcionar regularización suficiente sin ser excesivamente restrictivo.
    - Prevenir el sobreajuste mientras se mantiene la capacidad de aprendizaje.
    - Permitir ajuste fino de la regularización según la complejidad del modelo.
- **Parámetros fijos:** Además, se establecieron parámetros fijos basados en buenas prácticas:
    - *Attention dropout*: 0.1 para regularización específica del mecanismo de atención.
    - *Warmup steps*: 100 para estabilizar el entrenamiento inicial.
    - Número de épocas: 200 para permitir convergencia adecuada.
    - Pct *start*: 0.3 para optimizar el *scheduling* del *learning rate*.

Estos rangos se definieron específicamente para el contexto de diagnóstico médico, considerando tanto la complejidad de los datos como la necesidad de mantener la interpretabilidad del modelo.

4.1.4. Proceso de entrenamiento

Se implementaron diversas técnicas de optimización y regularización para asegurar un entrenamiento estable y eficiente:

- **Optimizador Adam:** Seleccionado por su capacidad de adaptación automática de las tasas de aprendizaje y su eficacia probada en modelos Transformer, configurado con:
    - *Learning rate* inicial optimizado para el problema.

- *Weight decay* para regularización adicional.
- *Warmup steps* para estabilización inicial.
- Pct *start* para optimizar el *scheduling* del *learning rate*.

- ***Early stopping*:** Implementado con una paciencia adaptada para prevenir el sobreajuste mientras se asegura suficiente tiempo para la convergencia del modelo.
- ***Dropout* diferenciado:** Implementado con tasas distintas para:
  - Características categóricas (*dropout* completo).
  - Características continuas (*dropout* reducido al 50 % del general).
  - Mecanismo de atención (*attention dropout*).
- ***Gradient clipping*:** Implementado para prevenir la explosión de gradientes, crucial en arquitecturas profundas como Transformer.

Algoritmo 2 Entrenamiento del modelo con early stopping

Input: Conjunto de datos de entrenamiento y validación, hiperparámetros optimizados, paciencia.
Output: Modelo entrenado con los mejores pesos.

```
modelo ← crear modelo(mejores hiperparámetros)
mejor val acc ← 0
sin mejora ← 0
for época = 1 to max épocas do
    for batch en train loader do
        loss ← entrenar batch(modelo, batch)
        actualizar modelo(modelo, loss)
    end for
    val acc ← evaluar modelo(modelo, val loader)
    if val acc >mejor val acc + min improvement then
        mejor val acc ← val acc
        sin mejora ← 0
        guardar mejor modelo(modelo)
    else
        sin mejora ← sin mejora + 1
    end if
    if sin mejora >= paciencia then
        break
    end if
end for
return cargar mejor modelo()
```

4.1.5. Proceso de optimización

La optimización del modelo se realizó en dos fases principales, representadas por los algoritmos 1 y 2. En la primera fase (Algoritmo 1), se utilizó un algoritmo genético para realizar una búsqueda exhaustiva de hiperparámetros, evaluando múltiples configuraciones mediante validación cruzada para encontrar la combinación óptima. Una vez identificados los mejores hiperparámetros, estos se utilizaron en la segunda fase (Algoritmo 2) para entrenar el modelo final, implementando técnicas como *early stopping* para garantizar la generalización del modelo. Esta aproximación en dos fases permitió no solo encontrar la mejor configuración del modelo, sino también asegurar un entrenamiento robusto y eficiente.

4.2. Experimentos

Durante el desarrollo se realizaron diversas mejoras incrementales, como se muestra en la Tabla 4.1. Cada implementación del modelo con diferentes configuraciones contribuyó a mejorar diferentes aspectos del rendimiento:

Tabla 4.1

Resultados de evaluación de los modelos en el dataset de validación.

Implementación	***Loss***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
01.- Implementación Base	0.3081	**0.8845**	0.8879	0.8803	0.8841
02.- Limpieza y Ajustes (A)	0.2971	0.8766	0.8744	0.8797	0.8770
02.- Limpieza y Ajustes (B)	0.2984	0.8769	0.8620	0.8979	0.8795
03.- Optimización de *Batch*	0.3004	0.8772	0.8596	**0.9016**	0.8802
04.- Optimización de Memoria	0.3088	0.8712	0.8544	0.8949	0.8742
05.- Optimización en Normalización	**0.2952**	0.8800	0.8734	0.8888	0.8811
06.- Optimización Final (A)	0.2961	0.8833	**0.8958**	0.8676	**0.8815**
06.- Optimización Final (B)	0.3240	0.8617	0.8695	0.8512	0.8603

Nota. Los valores en negritas representan el mejor resultado obtenido por métrica.

Como se puede observar en la tabla, algunas implementaciones (02.- Limpieza y Ajustes y 06.- Optimización Final) presentan múltiples resultados. Esto se debe a que, dentro del espacio de búsqueda de hiperparámetros definido (que incluye rangos para *Transformer_layers*, *Transformer_heads*, *embedding_dim*, *batch_size*, entre otros), se evaluaron diferentes configuraciones

para la misma implementación base del código. Por ejemplo, en la implementación 02, se probaron distintas combinaciones de estos hiperparámetros manteniendo la misma estructura de código optimizada, resultando en dos modelos con rendimientos diferentes:

- Primera configuración: *Loss*: 0.2971, *Accuracy*: 0.8766, *Precision*: 0.8744, *Recall*: 0.8797, *F1 Score*: 0.8770.
- Segunda configuración: *Loss*: 0.2984, *Accuracy*: 0.8769, *Precision*: 0.8620, *Recall*: 0.8979, *F1 Score*: 0.8795.

De manera similar, para la implementación 06 (Optimización Final), se exploraron dos configuraciones diferentes de hiperparámetros dentro de los rangos establecidos, llevando a resultados distintos:

- Primera configuración: *Loss*: 0.2961, *Accuracy*: 0.8833, *Precision*: 0.8958, *Recall*: 0.8676, *F1 Score*: 0.8815.
- Segunda configuración: *Loss*: 0.3240, *Accuracy*: 0.8617, *Precision*: 0.8695, *Recall*: 0.8512, *F1 Score*: 0.8603.

Esta exploración de diferentes configuraciones de hiperparámetros dentro de una misma implementación permitió una evaluación más exhaustiva del espacio de soluciones, contribuyendo a una mejor comprensión del comportamiento del modelo bajo distintas condiciones de entrenamiento.

4.2.1. Descripción de las implementaciones

- **01.- Implementación Base:** Primera versión funcional del modelo TabTransformer, estableciendo la arquitectura fundamental y los componentes principales.
- **02.- Limpieza y Ajustes:** Optimización del código mediante la eliminación de archivos innecesarios y mejoras en la definición del *batch*, resultando en una estructura más limpia y eficiente.
- **03.- Optimización de *Batch*:** Correcciones específicas en el manejo de *batches* para mejorar el rendimiento y la eficiencia del entrenamiento.

- **04.- Optimización de Memoria:** Implementación de mejoras significativas en la gestión de memoria, incluyendo la limpieza adecuada de recursos y optimización del uso de GPU.
- **05.- Optimización en Normalización:** Refinamiento del proceso de normalización de datos, mejorando la estabilidad del entrenamiento.
- **06.- Optimización Final:** Últimos ajustes y mejoras en la configuración general del modelo, buscando optimizar el rendimiento global.

4.2.2. Consideraciones de implementación

Durante el proceso de desarrollo, se identificaron y abordaron varios desafíos técnicos:

- **Gestión de memoria:** Se implementaron soluciones para evitar fugas de memoria y optimizar el uso de recursos GPU.
- **Procesamiento por lotes:** Se ajustó el tamaño y manejo de *batches* para balancear eficiencia y rendimiento.
- **Normalización de datos:** Se refinó el proceso de normalización para mejorar la convergencia del modelo.

4.2.3. Análisis estadístico de los resultados

Para fortalecer la validez de los resultados obtenidos, se aplicó una prueba ANOVA de un factor con el fin de determinar si existen diferencias estadísticamente significativas entre las métricas de desempeño de las distintas versiones del modelo TabTransformer (como *Accuracy* y *F1-score*).

Se evaluaron las ocho implementaciones, cada una con ajustes específicos. A través de un análisis estadístico descriptivo se compararon las métricas *Loss*, *Accuracy*, *Precision*, *Recall* y *F1-score*, cuyos valores se resumen en la Tabla 4.2. La métrica con mayor variabilidad fue *Loss* (CV = 3.22 %), mientras que *Accuracy* presentó la menor (CV = 0.83 %). La Figura 4.1 ilustra visualmente esta comparación.

La implementación denominada “03.- Optimización de *Batch*” obtuvo el mayor valor de *Recall* (0.9016), con un buen equilibrio en otras métricas, lo cual respalda su posterior elección como

versión final del modelo.

Figura 4.1

Comparación visual de las métricas principales entre implementaciones del modelo TabTransformer. La barra azul corresponde a la implementación con el mejor Recall ('03.- Optimización de Batch').

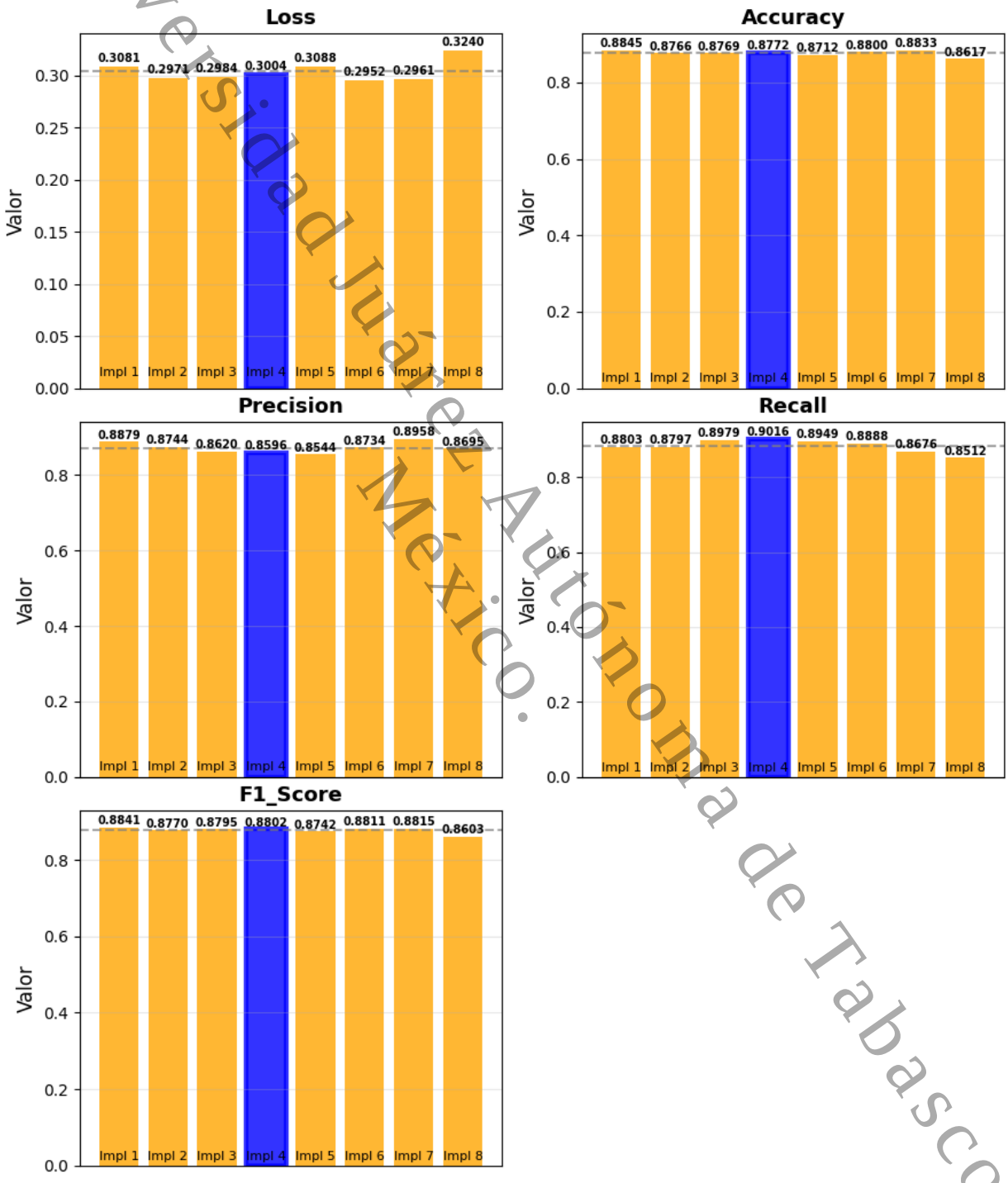


Tabla 4.2

Resumen estadístico de métricas por implementación.

Métrica	Media	Desv. Std.	Mínimo	Máximo	Rango	CV (%)
Loss	0.3035	0.0098	0.2952	0.3240	0.0288	3.22
Accuracy	0.8764	0.0073	0.8617	0.8845	0.0228	0.83
Precision	0.8721	0.0141	0.8544	0.8958	0.0414	1.62
Recall	0.8828	0.0169	0.8512	0.9016	0.0504	1.92
F1-score	0.8772	0.0075	0.8603	0.8841	0.0238	0.85

4.2.4. Selección del modelo final

Aunque varios modelos mostraron un rendimiento comparable en términos generales, se seleccionó como modelo final la implementación "03.- Optimización de *Batch*" que alcanzó el mejor *recall* (0.9016). Esta decisión se fundamenta en el contexto médico específico del diagnóstico de cardiopatía isquémica crónica, donde:

- Un falso negativo (no detectar la enfermedad cuando está presente) tiene consecuencias potencialmente más graves que un falso positivo.
- La capacidad de identificar correctamente los casos positivos (*recall*) es crucial para iniciar el tratamiento temprano.
- El costo de pruebas adicionales para confirmar un diagnóstico positivo es menor que el riesgo de no detectar la enfermedad.

Si bien otros modelos mostraron mejores resultados en *precision* o *F1-score*, la priorización del *recall* en este contexto médico específico justifica la selección de esta implementación como el modelo final para su despliegue en la práctica clínica.

4.3. Resultados

La evaluación del modelo se realizó utilizando un conjunto amplio de métricas, cada una elegida para analizar distintos aspectos de su rendimiento: la *precision* para medir la exactitud general del modelo, el *recall* para evaluar la capacidad de identificar correctamente los casos positivos

(crucial en diagnósticos médicos donde es preferible un falso positivo a un falso negativo), el *F1-score* para balancear *precision* y *recall*, y el AUC-ROC para evaluar la capacidad discriminativa del modelo a través de diferentes umbrales de clasificación. Esta combinación de métricas permite una evaluación holística del desempeño del modelo en el contexto específico del diagnóstico médico.

4.3.1. Configuración del entrenamiento

El modelo fue entrenado utilizando el optimizador Adam con una tasa de aprendizaje inicial de 0.0038. Se aplicó la función de pérdida de entropía cruzada binaria (BCELoss) dado que el problema es de clasificación binaria.

Se configuró el entrenamiento para un total de 200 épocas, con 100 pasos de calentamiento (*warmup steps*) y un porcentaje de inicio (pct *start*) de 0.3. Se empleó la estrategia de *early stopping*, estableciendo una paciencia de 50 épocas, lo que significa que si no había mejora en la precisión de validación dentro de este intervalo, el entrenamiento se detenía automáticamente para evitar el sobreajuste.

Además, se aplicó validación cruzada de 5 *folds* para evaluar el rendimiento del modelo en diferentes particiones del conjunto de datos y garantizar su capacidad de generalización.

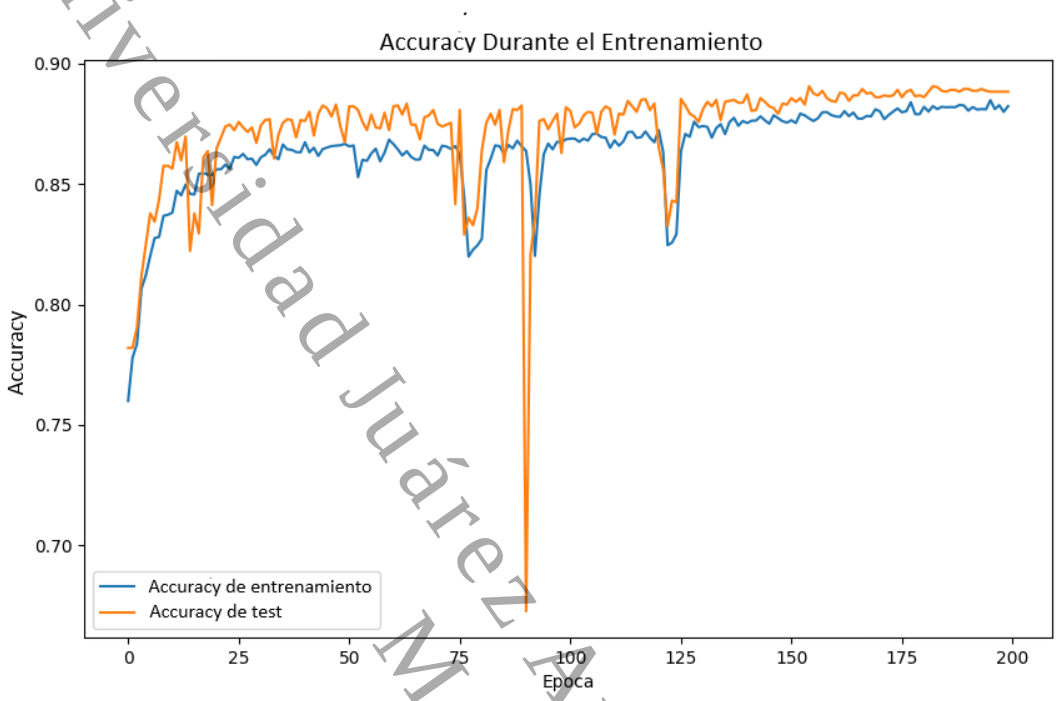
4.3.2. *Accuracy* durante el entrenamiento y la prueba

Se utilizó el *accuracy* para evaluar el rendimiento del modelo en cada época, ademas de presentar cómo mejora su capacidad para clasificar correctamente los datos, durante el proceso de entrenamiento.

La Figura 4.2 muestra cómo, a lo largo de las épocas, el *accuracy* del entrenamiento incrementó de manera constante al principio, hasta estabilizarse en las últimas iteraciones. Asimismo, el *accuracy* de validación siguió una tendencia similar, lo cual indicó que el modelo generaliza adecuadamente a los datos no vistos, aunque se pueden observar ligeras variaciones debido a la naturaleza de los datos de validación, que hacen uso de la validación cruzada, es por eso que la variación se observó al momento de pasar los datos de validación.

Figura 4.2

Accuracy durante entrenamiento y validación.



Nota. Elaboración propia con base en los registros del entrenamiento del modelo.

Como se muestra en la Tabla 4.3, el modelo alcanzó sus mejores métricas de prueba durante la época 155. Los valores reportados incluyen una *Loss* de 28.14, un *accuracy* de 89.06, una *precision* de 87.92, un *recall* de 90.83, y un *F1 Score* de 89.35. Esto muestra que el modelo tiene un buen equilibrio entre la precisión y la capacidad para identificar correctamente las clases, lo cual es crucial para problemas de clasificación. El valor de *F1 Score*, cercano al *recall*, sugiere una adecuada proporción entre falsos positivos y falsos negativos, lo cual es clave para obtener un rendimiento robusto en el contexto de la clasificación.

Tabla 4.3

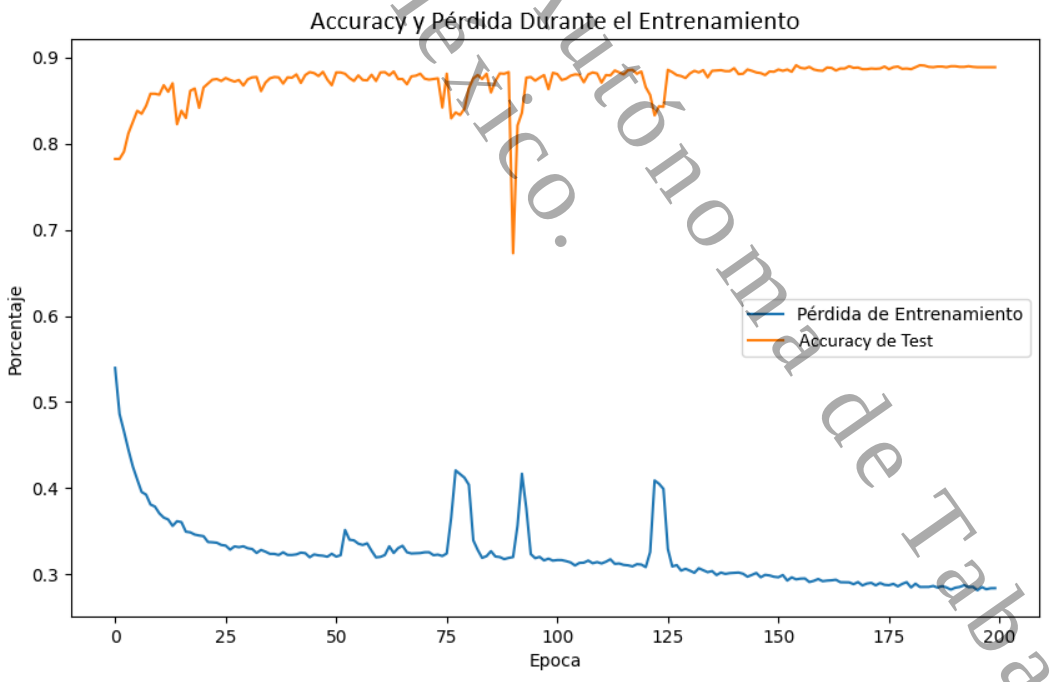
Métricas del mejor modelo en la fase de prueba (test).

Métrica	Valor
Loss	28.14 %
Accuracy	89.06 %
Precision	87.92 %
Recall	90.83 %
F1-score	89.35 %
Epoch	155

Se puede observar en la Figura 4.3, como la pérdida durante el entrenamiento disminuyó de manera constante mientras que el *accuracy* se estabilizó.

Figura 4.3

Evolución del accuracy y pérdida durante el entrenamiento del modelo.



Nota. Elaboración propia con base en los registros del entrenamiento del modelo.

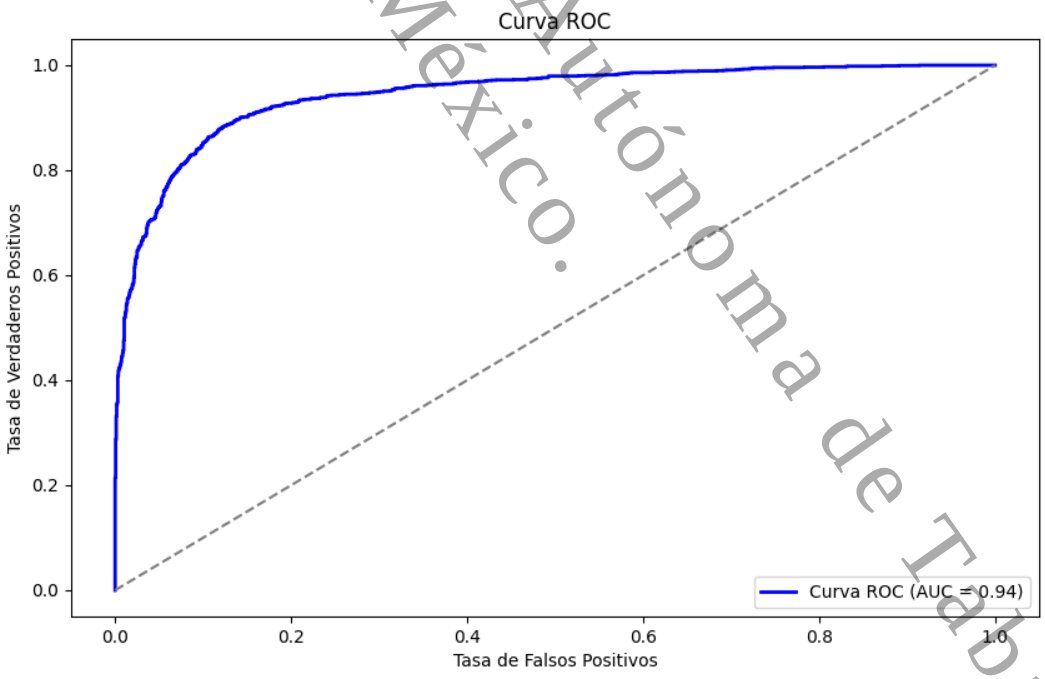
4.3.3. Rendimiento del modelo final

Para evaluar el rendimiento del modelo de red neuronal Transformer, se utilizaron diversas métricas de clasificación, tales como *accuracy, precision, recall, F1-score* y área bajo la curva ROC (AUC). Estas métricas permitieron cuantificar la capacidad de del modelo en la clasificación de pacientes con y sin cardiopatía isquémica crónica.

La Figura 4.4 muestra la curva ROC *(Receiver Operating Characteristic Curve)* obtenida durante el testeo del modelo. Se puede apreciar que el área bajo la curva (AUC) alcanzó un valor de 0.9434, lo cual sugiere un alto rendimiento del modelo en cuanto a su capacidad para diferenciar entre las clases positivas y negativas. Se muestra que la Curva ROC está bien trazada y se encuentra claramente por encima de la línea diagonal (que representa la decisión aleatoria).

Figura 4.4

Curva ROC y valor de AUC obtenido en el conjunto de test.



Nota. Elaboración propia a partir de las predicciones del modelo en el conjunto de prueba.

Como se muestra en la Tabla 4.4, los resultados indican un buen rendimiento del modelo, destacando un AUC de 0.9434.

Tabla 4.4

Resultados de la evaluación del modelo en el conjunto de prueba (test).

Métrica	Valor
Loss	30.04 %
Accuracy	87.72 %
Precision	85.98 %
Recall	90.16 %
F1 Score	88.02 %
Area Under the Curve	0.9434

4.3.4. Comparación con otros modelos

Para proporcionar una perspectiva más completa, se realizó una evaluación comparativa entre el modelo Transformer y otros modelos de clasificación, cuyos resultados se presentan en las siguientes tablas. Los resultados mostraron que el modelo Transformer ofrece una ventaja significativa en términos de *accuracy, precision, recall* y *F1-score*, tanto en la fase de validación (Tabla 4.5) como en la fase de test (Tabla 4.6). Esto justifica su selección como el modelo final para el diagnóstico de la cardiopatía isquémica crónica en pacientes con diabetes mellitus tipo 2.

Tabla 4.5

Comparación de métricas de rendimiento entre modelos en la fase de validación.

Métrica	Transformer	Regresión logística	KNN	Árboles de decisión
Accuracy	89.06 %	84.66 %	81.50 %	84.20 %
Precision	87.92 %	85.19 %	83.69 %	84.48 %
Recall	90.83 %	84.30 %	78.74 %	84.22 %
F1-score	89.35 %	84.74 %	81.15 %	84.35 %

Nota. Todos los resultados corresponden a la fase de validación (val).

Tabla 4.6

Comparación de métricas de rendimiento entre el mejor modelo Transformer en la fase de prueba (test), regresión logística (test), KNN (test) y árboles de decisión (test).

Métrica	Transformer	Regresión Logística	KNN	Árboles de decisión
Accuracy	87.72 %	83.74 %	80.67 %	83.26 %
Precision	85.98 %	83.88 %	82.30 %	82.92 %
Recall	90.16 %	83.54 %	78.19 %	83.78 %
F1-score	88.02 %	83.71 %	80.19 %	83.35 %

Nota. Elaboración propia con base en los resultados obtenidos en el conjunto de prueba (*test*).

4.3.5. Análisis de errores y limitaciones del modelo

Durante las diferentes pruebas y optimizaciones realizadas, se observó que el modelo mantuvo un rendimiento estable: el *accuracy* se mantuvo alrededor del 87-88 %, la *precision* entre 87-90, el *recall* entre 85-90 % y el *F1-score* cerca del 88 %. Esto sugiere que, a pesar de probar diferentes configuraciones y ajustes en el modelo, los resultados se mantuvieron en rangos similares.

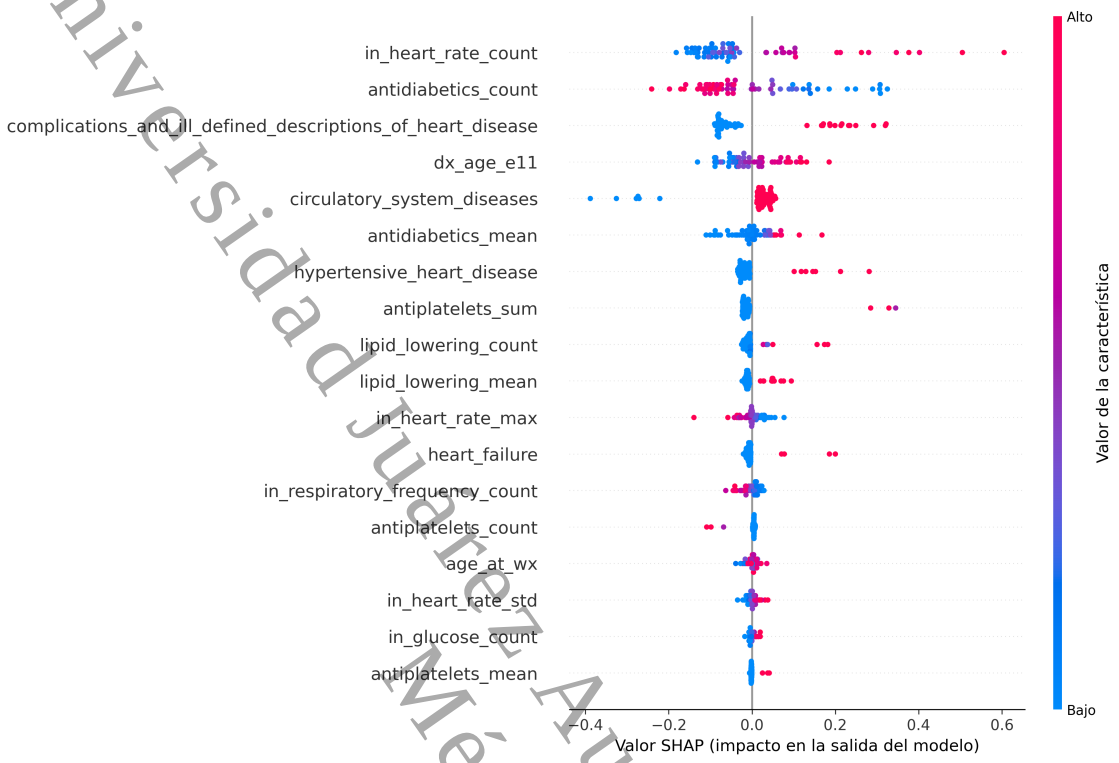
Estos resultados consistentes validan la efectividad del análisis exploratorio y el preprocesamiento de datos realizado previamente. Este tratamiento de datos permitió que el modelo mantuviera un rendimiento estable a través de las diferentes implementaciones y optimizaciones, así como en la comparación con otros modelos de *machine learning*, en los cuales los datos igual presentaron esa estabilidad al momento de hacer la comparativa.

4.3.6. Importancia de las características seleccionadas

Se realizó un análisis de importancia de características utilizando el método SHAP (*SHapley Additive Explanations*), que permitió validar la selección previa de variables realizada mediante los métodos de *Rank cor*, Chi-cuadrada, CFS y Correlación de Pearson. Los resultados del análisis SHAP confirmaron la relevancia de las variables seleccionadas, mostrando una consistencia particular con variables como `in_heart_rate_count, antidiabetics_count` y `complications_and_ill_defined_descriptions_of_heart_disease` las cuales también fueron identificadas como significativas en el proceso inicial de selección de características.

Figura 4.5

Importancia de las características según el método SHAP, validando la selección previa de variables.



Nota. Elaboración propia con base en el modelo final entrenado y el análisis SHAP.

Esto refuerza la robustez del proceso de selección de variables realizado en las etapas iniciales del estudio, confirmando la importancia de las 18 variables seleccionadas para el modelo final.

4.3.7. Factores de riesgo reconocidos en la medicina.

Como se mencionó en la sección Selección de características, el proceso de selección de características permitió identificar las variables más relevantes desde el punto de vista computacional, y el análisis SHAP nos muestra la importancia de esas características dentro del modelo de red neuronal, por otro lado la literatura médica reconoce un conjunto de factores de riesgo que determinan la cardiopatía isquémica crónica (FEC, 2024a; IMSS, 2009). Entre los principales se encuentran:

- Obesidad y sobrepeso.
- Sedentarismo.

- Dislipidemia (LDL elevado, HDL bajo, hipertrigliceridemia).
- Hipertensión arterial.
- Diabetes mellitus.
- Síndrome metabólico.
- Tabaquismo.
- Antecedentes familiares de cardiopatía isquémica.
- Edad avanzada.
- Sexo masculino (con aumento en mujeres tras la menopausia).

Sin embargo, varias de las variables destacadas computacionalmente no aparecen en la literatura médica clásica como factores de riesgo. Tal es el caso de `antiplatelets_count`, `antiplatelets_mean`, `lipid_lowering_mean`, `in_heart_rate_std` o `in_respiratory_frequency_count`. Estas variables, más que representar factores causales, parecen reflejar síntomas clínicos, mediciones del cuerpo o patrones de tratamiento con medicamentos que influyen en el diagnóstico. Este contraste evidencia que, mientras la literatura médica se centra en las causas de las enfermedades, el enfoque computacional puede capturar señales adicionales que complementan la práctica clínica y enriquecen la predicción del modelo. Conviene señalar además que, como se mencionó en la sección *Measurements*, variables como `fn_weight` (relacionada con obesidad/sobrepeso) estaban presentes en el dataset original, pero fue eliminada durante el proceso de rankeo a partir de implementar los distintos métodos de selección de características, lo que refuerza la diferencia entre la perspectiva clínica tradicional y el criterio de relevancia computacional.

4.3.8. Hiperparámetros del modelo TabTransformer

En la Tabla 4.7 se presentan los principales hiperparámetros utilizados en el modelo TabTransformer:

Tabla 4.7

Hiperparámetros del modelo TabTransformer.

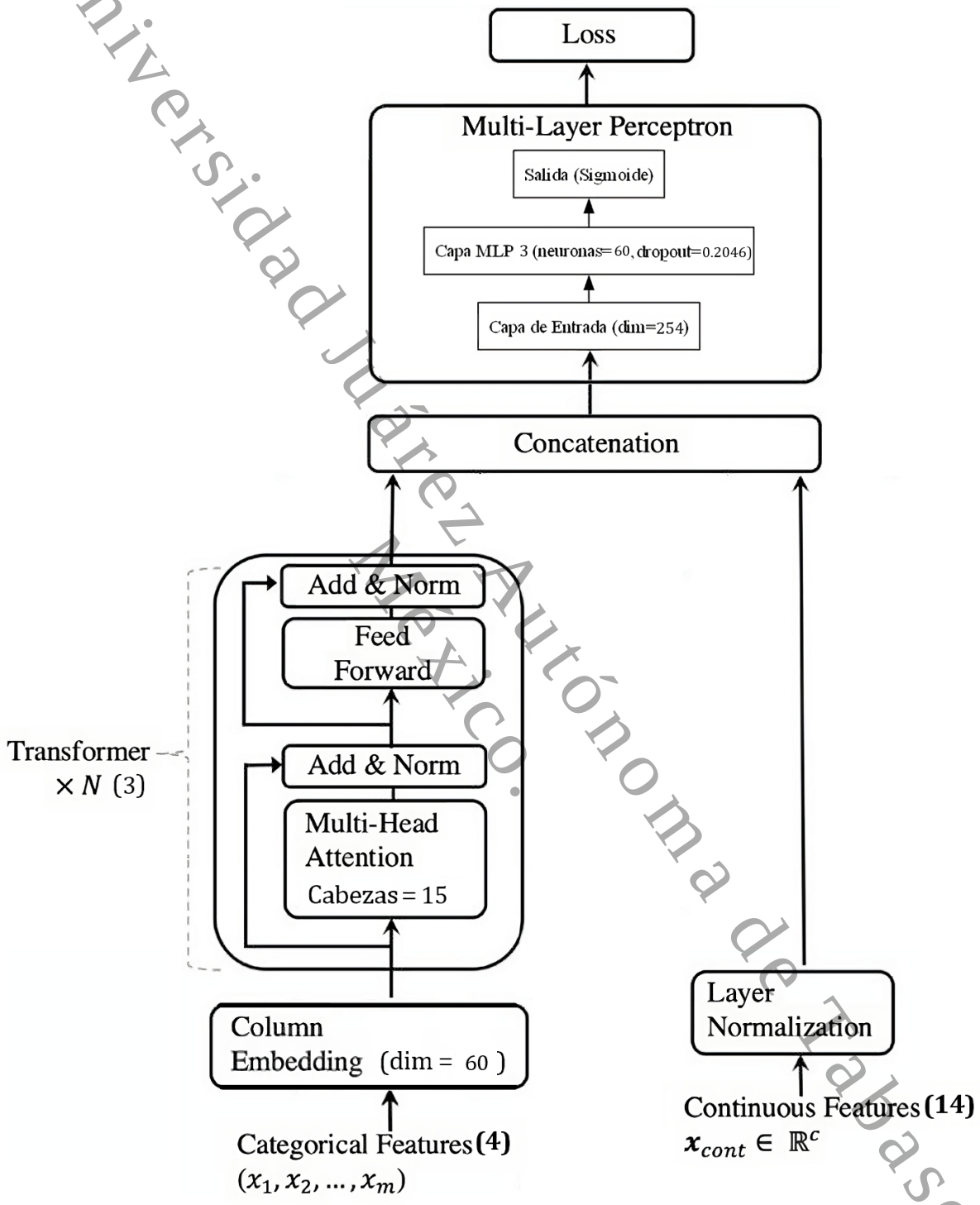
Hiperparámetro	Valor
embedding_dim	60
transformer_layers	3
transformer_heads	15
mlp_layers	3
dropout_rate	0.2046
attention_dropout	0.1
learning_rate	0.0038
weight_decay	0.0004
batch_size	36
num_epochs	200
warmup_steps	100
pct_start	0.3
categorical_features_count	4
continuous_features_count	14
input_dim	334

4.3.9. Diagrama de arquitectura del modelo TabTransformer

En la Figura 4.6, se presenta la arquitectura del modelo TabTransformer utilizado en el presente estudio:

Figura 4.6

La arquitectura del TabTransformer obtenida en esta investigación.



Nota. Elaboración propia. La arquitectura muestra las capas embebidas y la salida del MLP.

Capítulo 5

Discusión

Los resultados obtenidos en este estudio demuestran el potencial del modelo TabTransformer para el diagnóstico de la CIC en pacientes con DM2, superando significativamente los enfoques tradicionales de aprendizaje automático.

5.1. Interpretación de los resultados

El alto rendimiento del modelo TabTransformer puede atribuirse a varios factores:

- Su capacidad para capturar relaciones complejas entre variables clínicas mediante el mecanismo de atención multi-cabeza.
- La integración efectiva de variables categóricas y numéricas.
- La selección previa de variables relevantes, que proporcionó una base sólida para el modelo.

El alto valor de *recall* (90.16 %), que indica una excelente capacidad para identificar correctamente los casos positivos de CIC, es particularmente destacable. En el contexto clínico, esta métrica es especialmente relevante, ya que minimiza los falsos negativos (pacientes que se les diagnóstico como no enfermos, cuando en realidad si están enfermos), permitiendo así la detección y el tratamiento oportuno de la enfermedad.

5.2. Comparación con otros autores

Basándonos en la revisión de la literatura, podemos determinar que nuestros resultados son superiores a los de otros autores en términos de la proporción total de clasificaciones correctas en el diagnóstico de cardiopatía isquémica crónica. Logramos un *accuracy* del 87.72 %, mientras que un primer grupo de investigadores obtuvieron una precisión del 82 % (Silveri et al., 2020), y otros (Wang et al., 2023) alcanzaron un *accuracy* del 83.78 % en la clasificación de enfermedad coronaria. Es importante aclarar que los dos últimos estudios mencionados no se centraron en pacientes con DM2, ya que, a nuestro conocimiento, no existen estudios previos que utilicen una red neuronal basada en Transformer para diagnosticar la CIC en pacientes con DM2.

5.3. Limitaciones

Algunas de las limitaciones de este trabajo son las siguientes:

- Los resultados obtenidos dependen de los datos disponibles en el conjunto de datos `hk_database` de la colección DiabetIA, el cual, al ser una colección de datos capturados en escenarios del mundo real, presentaba valores faltantes. Esta situación llevó a la eliminación de algunos atributos que podrían haber influido en los resultados.
- Las características y la cantidad de datos disponibles en el conjunto de datos `hk_database` de la colección DiabetIA limitan la generalización de los resultados. La propuesta representa una herramienta de apoyo para la toma de decisiones, pero los resultados deben ser validados por un humano experto en el tema.
- Esta investigación solo utilizó el conjunto de datos `hk_database` de la colección DiabetIA, que contiene información clínica de pacientes mexicanos con DM2. Por lo tanto, el uso de otros conjuntos de datos podría llevar a variaciones en los resultados. Adicionalmente pudieran haber sesgos en los datos empleados en el sentido de que la población de enfermos que llevaron a la construcción del *dataset* pueden poseer un patrón de características; por ejemplo, hombres, casados, de mediana edad, con hijos, e ingresos bajos.

5.4. Implicaciones clínicas

Este estudio tiene importantes implicaciones para la práctica clínica:

- Podría ayudar a identificar pacientes de alto riesgo que requieren una evaluación cardiovascular más exhaustiva.
- El análisis de la importancia de las características proporciona información sobre los factores más relevantes para el diagnóstico de CIC, lo que podría orientar la práctica clínica.

La implementación de este tipo de modelo en sistemas de apoyo a la toma de decisiones clínicas podría contribuir significativamente a la detección temprana de la CIC en pacientes diabéticos, mejorando potencialmente el pronóstico y reduciendo la morbilidad y mortalidad asociadas.

Capítulo 6

Contribuciones, conclusiones y trabajos futuros

6.1. Respuestas a las preguntas de investigación

6.1.1. Estructura óptima de la red neuronal para el diagnóstico de cardiopatía isquémica crónica

En respuesta a la primera pregunta de investigación: "¿Cuál es la estructura de red neuronal artificial en cuanto a capas, neuronas por capa y funciones de activación que diagnostique la cardiopatía isquémica crónica con al menos el 80 % de exactitud?", esta investigación determinó que la estructura óptima es un modelo TabTransformer con las siguientes características:

- **Arquitectura general:** TabTransformer adaptado específicamente para datos tabulares médicos.
- ***Embedding dimension*:** 60.
- **Capas Transformer:** 3 capas.
- **Transformer *heads*:** 15 cabezas de atención por capa.
- **Capas MLP:** 3 capas para la clasificación final.

- ***Dropout rate*:** 0.2046 para regularización.
- ***Attention dropout*:** 0.1.
- **Características de entrada:** 18 variables seleccionadas (4 categóricas y 14 continuas).
- **Tamaño de *batch*:** 36.

Esta configuración superó ampliamente el umbral mínimo establecido del 80 % de exactitud, alcanzando un *accuracy* de 87.72 % en el conjunto de prueba. El modelo demostró además un rendimiento superior a otros enfoques tradicionales de aprendizaje automático como regresión logística (83.74 %), k-vecinos más cercanos (80.67 %) y árboles de decisión (83.26 %).

La robustez de este modelo quedó demostrada mediante validación cruzada de 5 *folds*, con una variación estándar mínima entre diferentes ejecuciones (±0.0076 en *precision*, ±0.0147 en *recall* y ±0.0071 en *F1 Score*), lo que confirma su estabilidad y capacidad de generalización.

6.1.2. Métricas de calidad para evaluación del modelo

En respuesta a la segunda pregunta de investigación: "¿Cuáles son las métricas de calidad más adecuadas para evaluar la precisión de la red neuronal artificial Transformer?", se determinó que:

El ***Recall* (Sensibilidad)** es la métrica más importante en este contexto clínico específico, basado en las siguientes razones:

- Mide la capacidad del modelo para identificar correctamente los casos positivos (pacientes con cardiopatía isquémica crónica).
- Minimiza los falsos negativos, que representan el error más crítico en contextos médicos donde no detectar la enfermedad cuando está presente puede tener consecuencias potencialmente graves.
- El modelo seleccionado logró un *recall* del 90.16 % en el conjunto de prueba, lo que significa que más del 90 % de los pacientes con la condición fueron correctamente identificados.

Adicionalmente, se determinó que un enfoque de evaluación integral debe incluir:

- **AUC-ROC:** El área bajo la curva ROC (0.9434) demostró la excelente capacidad discriminativa del modelo a través de diferentes umbrales de clasificación.
- ***F1-Score*:** Como medida equilibrada entre *precision* y *recall* (88.02 %), particularmente relevante en conjuntos de datos con posible desbalance de clases.
- ***Accuracy* (Exactitud):** Para medir el rendimiento general del modelo (87.72 %), en cuanto a clasificación de enfermos y no enfermos.
- ***Precision*:** Para evaluar específicamente la proporción de verdaderos positivos entre todos los casos clasificados como positivos (85.98 %).

La selección final del modelo (implementación "03.- Optimización de *Batch*") se basó prioritariamente en la maximización del *recall* (90.16 %), dado que en este contexto médico específico, el costo de pruebas adicionales para confirmar un diagnóstico positivo es considerablemente menor que las implicaciones de no detectar la enfermedad, priorizando así la seguridad del paciente.

6.2. Contribuciones principales

Las contribuciones principales de esta investigación son:

- Desarrollo de un modelo TabTransformer específicamente optimizado para el diagnóstico de cardiopatía isquémica crónica en pacientes con diabetes mellitus tipo 2, con un rendimiento superior a lo que ofrecen enfoques tradicionales.
- Identificación de las 18 variables más relevantes para el diagnóstico de CIC en pacientes con DM2, validadas mediante el método SHAP.
- Determinación de la importancia del *recall* como métrica prioritaria en el contexto del diagnóstico médico de la CIC.
- Implementación y evaluación de un algoritmo genético adaptativo para la optimización de hiperparámetros del modelo TabTransformer.
- Desarrollo de un *pipeline* completo de preprocesamiento de datos médicos, incluyendo:
    - Metodología robusta para el tratamiento de datos nulos en registros médicos.

    - Proceso de validación de *outliers* en el contexto médico.
    - Técnicas de balanceo de clases para datos médicos desbalanceados.
- Validación de la estabilidad y robustez del modelo a través de múltiples implementaciones.

6.3. Conclusiones generales

Los resultados de esta investigación demuestran que:

- El modelo TabTransformer es altamente efectivo para el diagnóstico de cardiopatía isquémica crónica en pacientes con diabetes mellitus tipo 2, superando significativamente los umbrales establecidos de 80 % de exactitud en estudios similares.
- La arquitectura optimizada mediante algoritmos genéticos adaptativos permitió alcanzar una configuración robusta y estable a través de múltiples ejecuciones.
- El mecanismo de atención multi-cabeza del Transformer permite capturar relaciones complejas entre variables clínicas, lo que se traduce en un diagnóstico más preciso.
- En el contexto del diagnóstico médico, el *recall* debe priorizarse sobre otras métricas para minimizar los falsos negativos y tratar de asegurar la detección de la enfermedad.
- La estabilidad en los resultados (87-88 % *accuracy*, 85-90 % *recall*) sugiere que se alcanzó un límite natural en el rendimiento dado el conjunto de datos disponible.
- El análisis exhaustivo de los datos demostró la importancia del preprocesamiento en datos médicos para obtener resultados confiables.
- La selección de características mediante múltiples métodos proporcionó un conjunto robusto de variables predictivas.

6.4. Trabajos futuros

Esta investigación abre diversas líneas para trabajos futuros. Visualizamos los siguientes:

- Validación del modelo con conjuntos de datos externos y más diversos para evaluar su generalización.
- Integración del modelo en sistemas de apoyo a la decisión clínica para su evaluación en entornos reales.
- Investigación sobre la combinación de datos clínicos con otras modalidades de datos, como imágenes médicas o datos genómicos, para mejorar aún más la precisión diagnóstica.
- Implementación de mecanismos de atención específicos para series temporales médicas.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Modelo Transformer para diagnosticar cardiopatía isquémica crónica derivada de diabetes mellitus tipo 2
Autor:	Orlando Flores Custodio
ORCID:	https://orcid.org/0009-0004-4492-6478
Resumen:	La cardiopatía isquémica crónica (CIC) constituye una complicación frecuente en pacientes con diabetes mellitus tipo 2 (DM2), y su diagnóstico oportuno resulta fundamental para mejorar el tratamiento y pronóstico del paciente. Sin embargo, el diagnóstico se dificulta por la presencia atípica de síntomas y la diversidad de factores de riesgo. Esta investigación presenta el desarrollo, implementación, optimización y evaluación de un modelo de red neuronal tipo Transformer adaptado para datos tabulares (TabTransformer), capaz de diagnosticar CIC a partir de registros clínicos reales de pacientes con DM2. Se utilizó un conjunto de datos con información longitudinal de pacientes mexicanos atendidos en el Instituto Mexicano del Seguro Social de Michoacán, México durante el periodo comprendido entre los años 2005 y 2020. El modelo alcanzó un rendimiento destacable con un *accuracy* de 87.72 %, un *recall* de 90.16 % y un valor de 0.9434 para el área bajo la curva ROC. Se priorizó especialmente el *recall* como métrica principal, aspecto crítico en el diagnóstico de enfermedades cardiovasculares. Mediante un proceso riguroso de selección de características e implementando un algoritmo genético adaptativo para la optimización de hiperparámetros, se identificaron 18 variables determinantes para el diagnóstico. El modelo TabTransformer superó significativamente a los enfoques tradicionales de aprendizaje automático como regresión logística, k-vecinos más cercanos y árboles de decisión, demostrando una mayor capacidad para capturar relaciones complejas entre variables clínicas gracias a su mecanismo de atención. Los resultados confirman la efectividad del enfoque propuesto para el diagnóstico de CIC en pacientes diabéticos.
Palabras clave:	Inteligencia Artificial, Redes neuronales, Aprendizaje automático.
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

Abeliuk, A., & Gutiérrez, C. (2021). Historia y evolución de la inteligencia artificial. *Revista Bits de Ciencia*, (21), 14-21.

Acharya, B. B., & Shaileshbhai, G. D. (2024). Comparative Analysis of Machine Learning Algorithms: KNN, SVM, Decision Tree and Logistic Regression. *International Journal for Research in Applied Science and Engineering Technology*, *12*.

AHA. (2024). *¿Qué es la presión arterial alta?* https://www.heart.org/-/media/files/health-topics/answers-by-heart/answers-by-heart-spanish/what-is-highbloodpressure_span.pdf

Alkhodari, M., Azman, S. K., Hadjileontiadis, L. J., & Khandoker, A. H. (2022). Ensemble transformer-based neural networks detect heart murmur in phonocardiogram recordings. *2022 Computing in Cardiology (CinC)*, *498*, 1-4.

Almanza, O., Chia, E., De la Cruz, A., Tello, T., & Ortiz, P. J. (2017). Frecuencia de factores asociados a hipoglicemia en el adulto mayor diabético admitido en el servicio de emergencia de un hospital nacional. *Revista Médica Herediana*, *28*(2), 93-100.

AWS. (2023). ¿Qué es una red neuronal? https://aws.amazon.com/es/what-is/neural-network/

Ayala, A. E. G. (2007). Trastornos de la temperatura corporal. *Offarm*, *26*(7).

Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., & Elmaghraby, A. (2020). Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information*, *11*(4), 207. https://www.mdpi.com/2078-2489/11/4/207

Basto-Abreu, A., López-Olmedo, N., Rojas-Martínez, R., Aguilar-Salinas, C. A., Moreno-Banda, G. L., Carnalla, M., Rivera, J. A., Romero-Martínez, M., Barquera, S., & Barrientos-Gutiérrez, T. (2023). Prevalencia de prediabetes y diabetes en México: Ensanut 2022. *Salud Pública de México*, *65*, s163-s168.

Blondin, N. A. (2014). Diagnosis and management of periodic hypothermia. *Neurology: Clinical Practice*, *4*(1), 26-33.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

Coursera. (2025). 10 Machine Learning Algorithms to Know in 2025. https://www.coursera.org/articles/machine-learning-algorithms

CUN. (2024). *Hiperglucemia: qué es, síntomas, diagnóstico y tratamiento. Clínica Universidad de Navarra*. https://www.cun.es/enfermedades-tratamientos/enfermedades/hiperglucemia

Díaz, J. (2019). La realidad de la Inteligencia Artificial en Salud. *Instituto de Ingeniería del Conocimiento*, *20*. https://www.iic.uam.es/lasalud/realidad-inteligencia-artificial-salud/

El Jerjawi, N. S., & Abu-Naser, S. S. (2018). Diabetes prediction using artificial neural network. *International Journal of Advanced Science and Technology*, *121*, 55-64.

Entralgo, P. L. (1981). Los orígenes del diagnóstico médico. *Dynamis: Acta Hispanica ad Medicinae Scientiarumque Historiam Illustrandam*, *1*, 3-15.

Espinar Lara, R. (2018). Modelos de clasificación con datos no balanceados.

Fan, P., Shu, Y., & Han, Y. (2022). Transformer embedded with learnable filters for heart murmur detection. *2022 Computing in Cardiology (CinC)*, *498*, 1-4.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

FEC. (2024a). Cardiopatía isquémica. https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatia-isquemica.html

FEC. (2024b). *Frecuencia cardiaca*. https://fundaciondelcorazon.com/prevencion/marcadores-de-riesgo/frecuencia-cardiaca.html

Galván, M., & Medina, F. (2007). Imputación de datos: teoría y práctica. *Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL)*, (No. 4755).

Gálvez, M. M. (2013). Algunos hitos históricos en el desarrollo del diagnóstico médico por imágenes. *Revista Médica Clínica Las Condes*, *24*(1), 5-13.

Garcia, A. G. (2023). *Base de Datos DiabetIA*. https://repositorio-salud.conacyt.mx/jspui/handle/1000/296

Han, S., & Xiao, L. (2022). An improved adaptive genetic algorithm. *SHS web of conferences*, *140*, 1044.

Hayes, D. D. (2005). Bradicardia: ¿frecuencia cardíaca lenta? ¡Actuar rápido! *Nursing (Ed. española)*, *23*(4), 26-33.

Hodelín Maynard, E. H., Maynard Bermúdez, R. E., Maynard Bermúdez, G. I., & Hodelín Carballo, H. (2018). Complicaciones crónicas de la diabetes mellitus tipo II en adultos mayores. *Revista Información Científica*, *97*(3), 528-537.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.

Huang, B., Jiang, Q., Wu, T., Shen, Q., Wang, W., Wang, S., Huang, Y., Wang, S., Huang, P., Lin, M., Shi, X., & Li, X. (2022). Hypoglycemia unawareness identified by continuous glucose monitoring system is frequent in outpatients with type 2 diabetes without receiving intensive therapeutic interventions. *Diabetology & Metabolic Syndrome*, *14*(1), 180.

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint arXiv:1904.05342.*

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

Ignacio, J. B. (2019). *Clasificación con datos desbalanceados | Aprende Machine Learning*. https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

IMSS. (2009). Guía de Referencia Rápida: Diagnóstico y Tratamiento de Cardiopatía Isquémica Crónica. https://www.imss.gob.mx/sites/all/statics/guiasclinicas/000GRRCardiopatiaIsquemica.pdf

Katar, O., & Yildirim, O. (2023). An explainable vision transformer model based white blood cells classification and localization. *Diagnostics*, *13*(14), 2459.

Kinha, D., Gulati, G., Baranwal, A., & Vishwakarma, D. K. (2023). Classifying Heart Sounds for Disease Detection Using Deep Learning Methods. *2023 4th International Conference for Emerging Technology (INCET)*, 1-6. https://doi.org/10.1109/INCET57972.2023.10170444

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137-1145.

Lázaro, M. R., Hernández, M. D. J. G., Cocotle, J. J. L., León, A. C., & de los Santos, R. R. (2022). Calidad de vida en pacientes con diabetes mellitus tipo 2 en Tabasco, México. *Revista Iberoamericana de Enfermería Comunitaria: RIdEC*, *15*(1), 24-31.

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, *85*, 102762.

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., & Rahimi, K. (2022). Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, *27*(2), 1106-1117.

Martin, B., & Sanz Molina, A. (2001). *Redes neuronales y sistemas difusos*.

Matas González, I. (2021). Clasificación de imágenes mediante redes neuronales convolucionales y técnicas de deep learning avanzadas: transformers. https://idus.us.es/handle/11441/126503

MedlinePlus. (2023). Diabetes tipo 2. https://medlineplus.gov/spanish/diabetestype2.html

MedlinePlus. (2024). Manejo de su glucemia: MedlinePlus enciclopedia médica. https://medlineplus.gov/spanish/ency/patientinstructions/000086.htm

Mendoza Romo, M. Á., Padrón Salas, A., Cossío Torres, P. E., & Orozco, M. S. (2018). Prevalencia mundial de la diabetes mellitus tipo 2 y su relación con el índice de desarrollo humano. *Revista Panamericana de Salud Pública*, *41*, e103.

Miragaya, M. A., & Magri, O. F. (2016). Ecuación más conveniente para predecir frecuencia cardíaca máxima esperada en esfuerzo. *Insuficiencia cardíaca*, *11*(2), 56-61.

Monajatipoor, M., Rouhsedaghat, M., Li, L. H., Jay Kuo, C. C., Chien, A., & Chang, K. W. (2022). Berthop: An effective vision-and-language model for chest X-ray disease diagnosis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 725-734.

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2023). Transformers in healthcare: a survey. *arXiv preprint arXiv:2307.00067.*

OMS. (2007). Tabla de índice de masa corporal. https://www.imss.gob.mx/sites/all/statics/salud/tablas_imc/admayor_imc.pdf

Pérez, J. A. R., & Vicuña, E. V. L. (2022). La obesidad como factor de riesgo asociado a diabetes mellitus tipo 2. *Ciencia Latina Revista Científica Multidisciplinar*, *6*(3), 296-322.

Piątkiewicz, P. (2016). Hypoglycemia in elderly type 2 diabetes patients. *Diabetes Manag*, *6*(3), 71-75.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Prechelt, L. (2002). Early stopping-but when?

qu4nt & scikit-learn. (2021). *6.4. Imputación de valores faltantes — documentación de scikit-learn - 0.24.1*. https://qu4nt.github.io/sklearn-doc-es/modules/impute.html

Rodríguez, L. M., & Mármol, L. (2017). Curso básico sobre hipertensión. Tema 4. Betabloqueantes. *Farmacia Profesional*, *31*(4), 20-25.

Rodríguez-Robles, N. A., & Mayek-Perez, N. (2022). Orientación y seguimiento en la alimentación, actividad física de pacientes diabéticos de Miguel-Alemán y Ciudad-Mier, Tamaulipas. *Revista-e Ibn Sina*, *13*(1), 1-14.

Roshanzamir, A., Aghajan, H., & Soleymani Baghshah, M. (2021). Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, *21*, 1-14.

Rowden, A. (2022). Tasa de respiración normal: Cómo medir en adultos y otras edades.

Sato, Y., Kakizawa, M., Aso, S. I., Takayama, M., Yamashita, K., Miyamoto, T., & Aizawa, T. (2020). Startling hyperglycaemia with transient beta cell stunning in a patient with type 2 diabetes. *Endocrine Journal*, *67*(1), 95-98.

Silveri, G., Merlo, M., Restivo, L., Sinagra, G., & Accardo, A. (2020). Novel Classification of Ischemic Heart Disease Using Artificial Neural Network. *2020 Computing in Cardiology*, 1-4.

Siptroth, J., Moskalenko, O., Krumbiegel, C., Ackermann, J., Koch, I., & Pospisil, H. (2023). Investigation of metabolic pathways from gut microbiome analyses regarding type 2 diabetes mellitus using artificial neural networks. *Discover artificial intelligence*, *3*(1), 19.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437.

Tripp, J., Santana-Quinteros, D., Perez-Estrada, R., Rodriguez-Moran, M. F., Arcos-Gonzalez, C., Mercado-Rios, J., Cristobal-Perez, F., Hernandez-Martinez, B. R., Nava-Aguilar, M. A., Gonzalez-Arroyo, G., Salazar-Fernandez, E. P., Quiroz-Armada, P. S., Cortes-Vieyra, R., Noriega-Cisneros, R., Zinzun-Ixta, G., Maldonado-Pichardo, M. C., Flores-Alvarez, L. J.,

Reyes-Granados, S. C., Chagolla-Morales, R., ... Lopez-Pineda, A. (2023). DiabetIA: Building Machine Learning Models for Type 2 Diabetes Complications. *medRxiv*.

Wang, K., Meng, H., & Wang, X. (2023). Application of Vision-Series Transformer in screening for coronary heart diseases using coronary CT angiography. *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, 421-425.

Yaque, P. M. (1988). *Clasificación de distribuciones y datos atípicos*. Universidad Complutense de Madrid (Spain).