UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

DISEÑO DE MODELOS EXPLICATIVOS DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE INSUFICIENCIA RENAL

TESIS PARA OBTENER EL GRADO DE:

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

SIMÓN JAVIER HERNÁNDEZ GASPAR

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

CUNDUACÁN, TABASCO, A: JUNIO 2026

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

DISEÑO DE MODELOS EXPLICATIVOS DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE INSUFICIENCIA RENAL

TESIS PARA OBTENER EL GRADO DE:

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

SIMÓN JAVIER HERNÁNDEZ GASPAR

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

CUNDUACÁN, TABASCO, A: JUNIO 2026

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día Veintidos del mes de Junio del año 2026, el que suscribe **Simón Javier Hernández Gaspar**, alumno del Programa de la **Doctorado en Ciencias de la Computación** con número de matrícula **231H18007**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Doctorado y titulada **Diseño de modelos explicativos de aprendizaje automático para la detección de insuficiencia renal** , dirigida por el Dr. Oscar Alberto Chávez Bosquez

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 22 de Junio de 2026.

Estudiante: Simon Javier Hernández Gaspar

Cunduacán, Tabasco, a 22 de junio de 2026
Oficio No. 1285/2026/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Simón Javier Hernández Gaspar
Egresado del Doctorado 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 **"Diseño de modelos explicativos de aprendizaje automático para la detección de insuficiencia renal"**, para presentar examen y obtener el Grado de Doctor en Ciencias de la Computación.

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

Atentamente

Dra. Laura Beatríz Vidal Turrubiates
Directora

C.c.p. Dr. Eduardo Hernández de la Cuz. - Encargado del despacho de la Coordinación de Posgrado.
Alumno.

DRA.*LBVT/EHC

Carta de Cesión de Derechos

Villahermosa, Tabasco a 22 de Junio de 2026.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Diseño de modelos explicativos de aprendizaje automático para la detección de insuficiencia renal** .

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: Simón Javier Hernández Gaspar

TESTIGOS

Dr. Oscar Alberto Chávez Bosquez

Dra. Maricela García Avalos

Índice general

Índice de tablas

Índice de figuras

Resumen

Esta tesis aborda la detección temprana de complicaciones renales en pacientes con diabetes tipo 2 mediante Inteligencia Artificial Explicable (XAI). El objetivo fue diseñar y analizar modelos de aprendizaje automático que no solo predigan el riesgo renal, sino que también identifiquen los factores clínicos asociados, comparando modelos interpretables con técnicas de explicación post-hoc.

En la primera fase, se entrenó el algoritmo Explainable Boosting Machine (EBM) con el dataset cardíaco de la Cleveland Clinic Foundation (UCI), obteniendo una precisión de 0.8525 con 14 variables clínicas y 303 instancias, lo que permitió interpretar globalmente factores como la edad y el tipo de dolor torácico. En la segunda fase, EBM se validó en el dataset UCI de enfermedad renal crónica, alcanzando una precisión del 99 % e identificando marcadores clave como la creatinina sérica y la gravedad específica.

En la tercera fase, se analizó el dataset DiabetIA (42,182 instancias de población mexicana), donde Random forest logró un rendimiento del 80 %; mediante LIME se identificó que la hipertensión esencial y la edad son los factores de riesgo dominantes. Finalmente, una auditoría cruzada entre EBM y SHAP reveló un sesgo crítico: la subestimación del riesgo en pacientes normotensos, donde la ausencia de hipertensión puede enmascarar otras señales de alarma.

Se concluye que XAI no solo iguala el rendimiento de los modelos de caja negra, sino que constituye una herramienta de auditoría clínica clave para reducir omisiones diagnósticas y fortalecer la medicina de precisión en el sistema de salud mexicano.

Palabras clave: Inteligencia Artificial Explicable, EBM, SHAP, LIME, Nefropatía Diabética, Auditoría Clínica.

Abstract

This thesis addresses the early detection of renal complications in patients with type 2 diabetes through Explainable Artificial Intelligence (XAI). The objective was to design and analyze machine learning models capable of not only predicting renal risk, but also identifying associated clinical factors by comparing inherently interpretable models with post-hoc explanation techniques.

In the first phase, the Explainable Boosting Machine (EBM) algorithm was trained on the Cleveland Clinic Foundation (UCI) heart disease dataset, achieving an accuracy of 0.8525 with 14 clinical variables and 303 instances, enabling global interpretation of factors such as age and chest pain type. In the second phase, EBM was validated on the UCI chronic kidney disease dataset, reaching 99 % accuracy and identifying key biomarkers such as serum creatinine and specific gravity.

In the third phase, the DiabetIA dataset (42,182 instances from a Mexican population) was analyzed, where Random Forest achieved 80 % performance; using LIME, essential hypertension and age were identified as the dominant risk factors. Finally, a cross-interpretability audit comparing EBM and SHAP revealed a critical bias: the underestimation of risk in normotensive patients, where the absence of hypertension may mask other warning signs.

It is concluded that XAI not only matches the performance of black-box models, but serves as a key clinical auditing tool to reduce diagnostic omissions and strengthen precision medicine within the Mexican healthcare system.

Keywords: Explainable Artificial Intelligence, EBM, SHAP, LIME, Diabetic Nephropathy, Clinical Audit.

Capítulo 1

Protocolo de tesis

1.1. Introducción

1.1.1. Inteligencia Artificial

La Inteligencia Artificial se define como el campo de la ciencia de la computación dedicado al desarrollo de sistemas capaces de emular capacidades cognitivas propias del ser humano. Esta disciplina abarca el diseño de mecanismos para la adquisición de conocimientos, la aplicación de la lógica para la resolución de problemas y la optimización continua de sus funciones a partir de la experiencia (Russell y Norvig, 2022).

El Aprendizaje automático es una rama de la inteligencia artificial que permite a un sistema aprender a partir de datos, en lugar de depender de una programación explícita (Mitchell, 2019). Sin embargo, no se trata de un proceso sencillo: conforme el algoritmo procesa datos de entrenamiento, es capaz de generar modelos cada vez más precisos. Un modelo de Aprendizaje automático es el resultado que se obtiene al entrenar un algoritmo con dichos datos; una vez entrenado, el modelo recibe una entrada y produce una salida correspondiente. Las técnicas de Aprendizaje automático son fundamentales para mejorar la precisión de los modelos predictivos. Dependiendo de la naturaleza del problema y del tipo y volumen de datos disponibles, existen distintos enfoques. Las principales categorías del aprendizaje automático son:

- Aprendizaje supervisado es un enfoque del Aprendizaje automático en el que el modelo se entrena utilizando un conjunto de datos previamente etiquetados. Cada elemento de

entrada está asociado a una respuesta correcta conocida (variable objetivo), lo que permite al algoritmo aprender una función de mapeo para predecir los resultados de nuevos datos no observados (Goodfellow et al., 2016).

- Aprendizaje no supervisado a diferencia del enfoque supervisado, esta metodología trabaja con datos que no poseen etiquetas ni respuestas predefinidas. Su propósito fundamental es explorar la estructura intrínseca de la información para descubrir patrones ocultos, correlaciones o agrupaciones naturales (clustering) sin la guía explícita de un supervisor (Hastie et al., 2009).

- Aprendizaje por refuerzo es un paradigma de aprendizaje basado en la interacción de un agente dinámico con su entorno. El sistema aprende a tomar decisiones secuenciales mediante un mecanismo de prueba y error, donde cada acción ejecutada recibe una penalización o una recompensa, con el objetivo final de maximizar el beneficio acumulado a largo plazo (Sutton y Barto, 2018).

La selección de un paradigma de Aprendizaje automático implica un compromiso inherente entre la capacidad predictiva del modelo y su nivel de interpretabilidad. En la Figura 1.1 se ilustra esta relación inversa (conocida como *trade-off*), donde se observa que los modelos con mayor precisión y complejidad estructural, tales como las redes neuronales multicapa o los métodos de ensamble (*boosting* y *bagging*), operan frecuentemente como "cajas negras", ofreciendo nula visibilidad sobre su proceso interno de toma de decisiones. Por el contrario, algoritmos estadísticos más tradicionales como la regresión lineal o los Árboles de decisión (*Decision trees*), aunque presentan limitantes en escenarios de alta dimensionalidad o relaciones no lineales, proveen un alto grado de explicabilidad intrínseca. Estas limitantes generan la necesidad de desarrollar metodologías que permitan mitigar dicha disparidad, dando origen al campo de estudio que se detalla a continuación.

1.1.2. Inteligencia Artificial Explicable

Gunning, 2016 introdujo el término XAI (*Explainable Artificial Intelligence*) como parte de un programa de la agencia DARPA, el cual estuvo dirigido a construir sistemas de IA capaces de

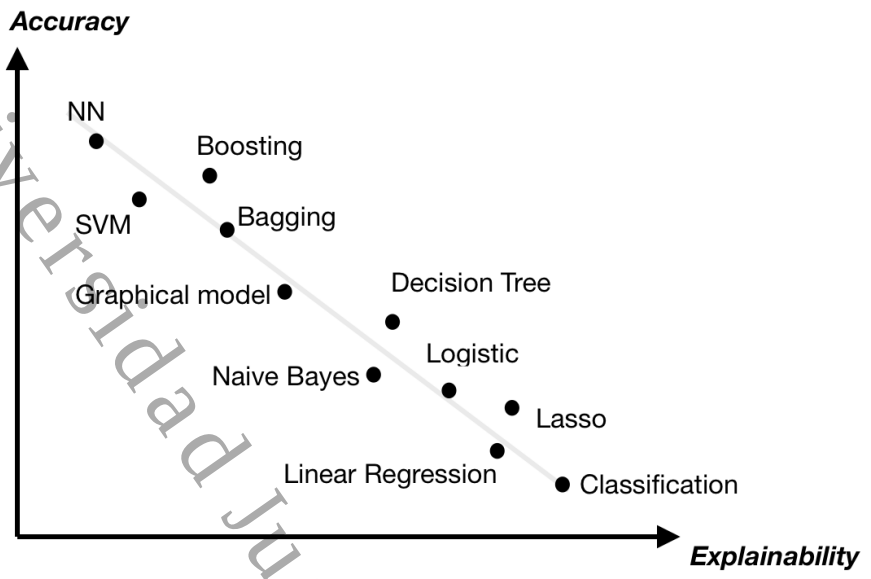

Figura 1.1. Comparación de algoritmos de explicativos contra precisión (Duval, 2019).

explicar su razón de ser.

Bajo este concepto, se propusieron diversas interrogantes que un usuario debería ser capaz de responder a partir de la explicación provista por el sistema de IA. Estas preguntas son (Gunning, 2016):

- ¿Por qué hiciste eso? (*Why did you do that?*)
- ¿Por qué no otra cosa? (*Why not something else?*)
- ¿Cuándo lo logras? (*When do you succeed?*)
- ¿Cuándo fallas? (*When do you fail?*)
- ¿Cuándo puedo confiar en ti? (*When can I trust you?*)
- ¿Cómo corrijo un error? (*How do I correct an error?*)

En este contexto, la IA explicable se define como un conjunto de procesos y métodos que permite a los usuarios comprender y confiar en los resultados generados por los algoritmos de IA y Aprendizaje automático (ML). Las explicaciones que acompañan a estos modelos pueden estar dirigidas a usuarios finales, operadores o desarrolladores, con la intención de abordar desafíos que van desde la adopción tecnológica hasta la gobernanza y el desarrollo seguro de sistemas.

Por consiguiente, un modelo de IA explicable es aquel que posee características o propiedades intrínsecas que facilitan la transparencia, la comprensibilidad y la capacidad de auditar o cuestionar las predicciones u outputs del sistema.

1.2. Marco Conceptual

1.2.1. Árboles de decisión

Los Árboles de decisión son representaciones gráficas que modelan secuencias de decisiones basadas en condiciones específicas. Este enfoque constituye uno de los algoritmos de aprendizaje supervisado más utilizados en el Aprendizaje automático, siendo capaz de resolver tareas tanto de clasificación como de regresión (Breiman et al., 1984).

Estructuralmente, un Árbol de decisión se compone de nodos y aristas. El proceso comienza en un nodo principal denominado raíz. A partir de este, el conjunto de datos se subdivide de forma jerárquica a través de bifurcaciones (ramas o aristas) basadas en los atributos de entrada. Cada nodo interno representa una regla de decisión sobre una característica, y el proceso de división se repite de manera recursiva hasta alcanzar los nodos terminales u hojas, los cuales representan la predicción final o la clase asignada (por ejemplo: verdadero/falso, comprar/vender).

Para construir el árbol óptimo y determinar la mejor partición en el nodo raíz y los subsiguientes, el algoritmo debe evaluar la calidad de las predicciones potenciales. Con el propósito de medir y valorar estas subdivisiones, se emplean funciones de impureza, siendo las más implementadas el “Índice de Gini” y la “Ganancia de información” (la cual hace uso de la entropía). Este proceso de partición continua se detiene cuando se alcanza la profundidad máxima configurada para el árbol, o bien, cuando los nodos contienen una cantidad mínima de muestras predefinida.

Índice de Gini

En el contexto del aprendizaje automático, el índice de Gini es una métrica utilizada para evaluar la impureza o diversidad de un nodo. Mide la probabilidad de que un elemento del conjunto de datos seleccionado al azar sea clasificado incorrectamente si se etiquetara aleatoriamente de acuerdo con la distribución de clases en ese nodo específico.

El valor del índice de Gini oscila entre 0 y 0.5 (en problemas de clasificación binaria), donde 0 representa una pureza perfecta (todas las muestras del nodo pertenecen a una única clase) y valores cercanos a 0.5 indican una máxima impureza (las clases están distribuidas de manera uniforme). Matemáticamente, se calcula mediante la siguiente ecuación:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

Tabla 1.1. Definición de variables para el cálculo del Índice de Gini.

Símbolo	Descripción
G	Coeficiente o Índice de Gini del nodo.
C	Número total de clases presentes en el problema.
p_i	Proporción de muestras que pertenecen a la clase i dentro del nodo.

Entropía

En el contexto de la teoría de la información de Shannon, la entropía es una métrica que cuantifica el grado de incertidumbre o desorden asociado a una variable aleatoria. Aplicada a los Árboles de decisión, la entropía mide la impureza de las clases en un nodo determinado. Si un nodo es completamente homogéneo (todas las muestras pertenecen a la misma clase), la entropía es igual a 0; por el contrario, si las clases están distribuidas uniformemente, la entropía alcanza su valor máximo de 1.

Formalmente, si un nodo contiene muestras pertenecientes a C clases distintas, y la probabilidad (o proporción) de que una muestra pertenezca a la clase i está dada por p_i, la entropía H se define mediante la suma ponderada de la cantidad de información:

$$H = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Ganancia de información

La Ganancia de información (*Information gain*) representa la reducción neta en la entropía que se logra al realizar una partición del conjunto de datos basada en un atributo específico. En

esencia, indica cuánta información aporta una característica para resolver la incertidumbre sobre la variable objetivo.

Definition 1. *Sea T un conjunto de entrenamiento donde cada elemento tiene la forma $(\mathbf{x}, y)$, siendo $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ el vector de características e y la etiqueta de clase correspondiente. Para un atributo específico a, se denota como $vals(a)$ al conjunto de todos los valores posibles que puede tomar dicho atributo. Así, el subconjunto de muestras en T para el cual el atributo a toma el valor v se define como:*

$$S_a(v) = \{\mathbf{x} \in T \mid x_a = v\}$$

Por consiguiente, la ganancia de información del atributo a respecto al conjunto de datos T se calcula restando la entropía ponderada de los subconjuntos resultantes a la entropía original del nodo:

$$IG(T, a) = H(T) - \sum_{v \in vals(a)} \frac{|S_a(v)|}{|T|} H(S_a(v))$$

A nivel metodológico, este enfoque ofrece ventajas significativas dentro del Aprendizaje automático, destacando por su transparencia y facilidad de interpretación, cualidades que reducen la opacidad del modelo y permiten una auditoría directa de sus reglas de decisión. Asimismo, es un algoritmo eficiente que requiere un preprocesamiento de datos mínimo en comparación con arquitecturas complejas. No obstante, los Árboles de decisión presentan desventajas críticas asociadas a su estabilidad; pequeños cambios en el conjunto de entrenamiento pueden derivar en estructuras arbóreas completamente distintas. Adicionalmente, son altamente susceptibles al sobreajuste (*overfitting*) cuando crecen sin restricciones, memorizando el ruido de los datos de entrenamiento y perdiendo capacidad de generalización ante muestras no observadas.

Para disminuir la complejidad y el sobreajuste, se implementa la técnica de poda (*pruning*), la cual consiste en remover de forma sistemática aquellas ramificaciones de niveles inferiores que no aportan un incremento significativo a la precisión predictiva del modelo. La poda permite simplificar la arquitectura del árbol, optimizando tanto su interpretabilidad como su rendimiento general. Una de las estrategias para regular este proceso consiste en establecer una diferencia

máxima en el riesgo basada en errores estándar. Esta regla permite al algoritmo seleccionar un subárbol más parsimonioso o simple cuya estimación del riesgo se mantenga estadísticamente próxima a la del subárbol con el menor riesgo absoluto, admitiendo un margen de tolerancia controlado en favor de la robustez y la simplicidad estructural del modelo.

1.2.2. Bosques aleatorios

El algoritmo de Bosques aleatorios (Random forest) constituye uno de los métodos de ensamble (*Ensemble learning*) más robustos y utilizados en el Aprendizaje automático (Breiman, 2001). Se fundamenta en la combinación de múltiples árboles de decisión individuales, donde la predicción final del modelo se determina mediante un mecanismo de votación mayoritaria para tareas de clasificación, o a través del promedio de las salidas en problemas de regresión. A diferencia de un árbol de decisión convencional, el cual genera reglas a partir de todo el conjunto de datos y características, Random forest introduce una doble aleatoriedad: selecciona subconjuntos aleatorios de muestras mediante la técnica de muestreo con reemplazo o *bagging*, y restringe la evaluación de divisiones en cada nodo a un subconjunto aleatorio de características (*feature bagging*). Este procedimiento garantiza que los árboles individuales presenten una baja correlación entre sí, lo cual permite capturar señales patrones informativos disímiles y maximizar el poder predictivo global del bosque.

Metodológicamente, este algoritmo destaca por su alta precisión predictiva y su notable flexibilidad para procesar grandes volúmenes de datos con cientos de variables de entrada sin requerir una reducción previa de dimensionalidad. Asimismo, posee una alta tolerancia al ruido y proporciona métricas intrínsecas para evaluar la importancia de las características (*feature importance*), así como mecanismos eficaces para disminuir el impacto de datos faltantes.

A pesar de estas ventajas, los bosques aleatorios presentan limitaciones críticas relacionadas con su interpretabilidad. Al integrar decenas o cientos de Árboles de decisión, el modelo se transforma en una estructura de “caja negra”, imposibilitando la inspección visual directa de sus reglas de decisión y justificando la necesidad de recurrir a técnicas de inteligencia artificial explicable. Adicionalmente, aunque es menos susceptible al sobreajuste que un árbol individual, Random forest tiende a sobreajustar en conjuntos de datos con ruido excesivo. Finalmente, el algoritmo

manifiesta un sesgo inherente al evaluar variables categóricas con un elevado número de niveles o al interactuar con grupos de características altamente correlacionadas, lo que puede distorsionar las estimaciones de importancia de las variables a menos que se apliquen correcciones específicas como las permutaciones parciales.

1.2.3. Máquinas de vectores de soporte (SVM)

Las Máquinas de vectores de soporte (*Support Vector Machines*, SVM) constituyen un paradigma de aprendizaje automático supervisado diseñado primordialmente para tareas de clasificación, aunque extensible a problemas de regresión (Vapnik, 1995). El objetivo fundamental de este algoritmo consiste en encontrar un hiperplano óptimo de separación en un espacio multidimensional que maximice el margen de distancia entre las clases del conjunto de datos. Aquellos puntos de entrenamiento que se sitúan en los límites de este margen y que definen la posición y orientación del hiperplano se denominan vectores de soporte. Dependiendo de la naturaleza de los datos, el margen puede configurarse como un margen rígido (*hard margin*), el cual exige una separación estricta y perfecta de las clases, o un margen flexible (*soft margin*), que tolera cierto grado de error para evitar el sobreajuste en distribuciones ruidosas.

Matemáticamente, la evaluación de la pertenencia de una muestra a una clase determinada se fundamenta en la geometría analítica y el cálculo vectorial. Dado un hiperplano definido por un vector de pesos $\mathbf{w}$ y un sesgo b, la clasificación de un vector de características de entrada $\mathbf{x}$ se determina mediante el signo del producto punto ponderado:

$$f(\mathbf{x}) = \text{signo}(\mathbf{w} \cdot \mathbf{x} + b)$$

Si el resultado de esta operación escalar es mayor a cero, la muestra se asigna a la clase positiva; de lo contrario, se clasifica en la clase negativa.

En escenarios donde los datos presentan una separación lineal perfecta, el algoritmo opera directamente en el espacio original. No obstante, las aplicaciones del mundo real frecuentemente exhiben fronteras de decisión no lineales. Para resolver esta limitante, SVM implementa el denominado “truco del kernel” (*kernel trick*), un mecanismo matemático que proyecta los datos originales a un espacio de características de mayor dimensión donde las clases se vuelven lineal-

mente separables, todo ello sin la necesidad de calcular explícitamente las coordenadas en dicho espacio de alta dimensionalidad.

Si bien algoritmos como Random forest y las SVM ofrecen una alta capacidad predictiva en datos complejos, su estructura interna —basada en ensambles de árboles o transformaciones de *kernels* no lineales— actúa como una caja negra, dificultando la interpretación directa de sus decisiones y requiriendo el uso de metodologías de explicabilidad.

1.2.4. Métodos y modelos de explicabilidad

Para implementar los conceptos de Inteligencia Artificial Explicable (XAI), en este trabajo se distinguen dos enfoques principales según el momento en que se genera la interpretación: los modelos intrínsecamente interpretables y los métodos de explicabilidad *post-hoc*.

Los modelos intrínsecamente interpretables presentan una estructura transparente por diseño, permitiendo comprender el proceso de decisión sin necesidad de herramientas externas. Dentro de esta categoría se empleará la *Explainable Boosting Machine* (EBM), un modelo basado en los Modelos Aditivos Generalizados (GAM) con componentes de interacción (Caruana et al., 2015). A diferencia de los métodos de *boosting* tradicionales, EBM entrena un árbol de decisión (o ensamble de árboles pequeños) sobre cada característica de forma independiente, y posteriormente sobre interacciones por pares. Matemáticamente, la predicción de un modelo EBM se expresa como una suma ponderada de funciones puntuación univariadas y bivariadas:

$$g(E[y]) = \beta_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{j \neq k} f_{j,k}(x_j, x_k)$$

Donde g es la función de enlace (*link function*), β_0 representa el sesgo o término de intersección, $f_i(x_i)$ es la función de puntuación que cuantifica el impacto de la característica individual x_i, y $f_{j,k}(x_j, x_k)$ modela los efectos de interacción entre pares de variables. Esta formulación permite que la contribución de cada variable sea directamente almacenable en tablas, visualizable y cuantificable, manteniendo un rendimiento predictivo comparable al de los modelos de caja negra.

Por otra parte, los métodos de explicabilidad *post-hoc* se aplican una vez que el modelo base (como SVM o Random forest) ha sido completamente entrenado, permitiendo extraer conocimiento de su comportamiento sin alterar su arquitectura interna. Para disminuir la opacidad de dichos

modelos, se emplearán técnicas agnósticas al modelo, destacando en primer lugar el método de Explicaciones Locales Interpretables Agnósticas al Modelo (*Local Interpretable Model-agnostic Explanations*, LIME) (Ribeiro et al., 2016). LIME busca explicar predicciones individuales aproximando localmente el comportamiento del modelo complejo mediante un modelo sustituto lineal (*surrogate model*) en el entorno de la muestra de interés. El objetivo de optimización de LIME se fundamenta en la siguiente expresión:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Donde f representa el modelo de caja negra original, g es el modelo sustituto explicable (perteneciente a la familia de modelos legibles G), $\pi_x(z)$ define la medida de proximidad o vecindad entre la muestra local x y las perturbaciones generadas z, $\mathcal{L}$ es la función de pérdida que evalúa la infidelidad del modelo g al aproximar a f en dicha vecindad, y $\Omega(g)$ cuantifica la complejidad del modelo sustituto para asegurar su interpretabilidad humana.

Como segundo método *post-hoc*, se utilizarán las Explicaciones Aditivas de Shapley (*SHapley Additive exPlanations*, SHAP) (Lundberg y Lee, 2017b). se emplea el marco de trabajo SHAP. Este método se basa en la teoría de juegos cooperativos para asignar de manera justa la contribución de cada variable clínica al resultado de la predicción final.

SHAP aproxima la predicción del modelo mediante un modelo de explicación aditivo, definido matemáticamente como:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{1.1}$$

Donde:

- g es el modelo explicativo.
- $z' \in \{0,1\}^M$ es el vector de características simplificadas (indicando si una característica está presente o ausente).
- M es el número de características de entrada.
- ϕ_0 es el valor base (*base value*), el cual representa la salida esperada del modelo sobre

todo el conjunto de entrenamiento ($E[f(x)]$).

- ϕ_j es la contribución individual de la característica j (el valor de Shapley).

Los valores de Shapley (ϕ_j) se calculan mediante la siguiente ecuación:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_x(S \cup \{j\}) - f_x(S) \right] \tag{1.2}$$

Donde:

- F es el conjunto de todas las características.
- S es un subconjunto de características que no contiene a la característica j.
- $f_x(S)$ es la predicción del modelo condicionado a los valores de las características en S.
- La diferencia $f_x(S \cup \{j\}) - f_x(S)$ representa el impacto marginal de añadir la característica j al subconjunto S.

Esta formulación matemática establece que la predicción final para un paciente específico, $f(x)$, es estrictamente la suma del riesgo base promedio más la contribución marginal de cada característica y sus interacciones:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{1.3}$$

La combinación de los enfoques presentados permite una evaluación integral del fenómeno en estudio. Por un lado, el uso de EBM proporciona una base de referencia transparente desde su construcción. Por otro lado, la aplicación de SHAP y LIME sobre modelos de alta complejidad (SVM y Random forest) asegura que la precisión predictiva no sacrifique la capacidad de auditoría clínica. Esta metodología de validación cruzada entre modelos intrínsecos y métodos *post-hoc* garantiza que los hallazgos no solo sean estadísticamente significativos, sino también interpretables y confiables para el personal médico.

1.3. Estado del arte

La bibliografia o revision literaria hasta el momento nos ha llevado entender como la IA puede ser explicable, y esto ocurrre bajo dos divisiones o áreas: modelos transparentes y modelos *post-hoc*. En la Figura 1.2 se describen ambas categorías.

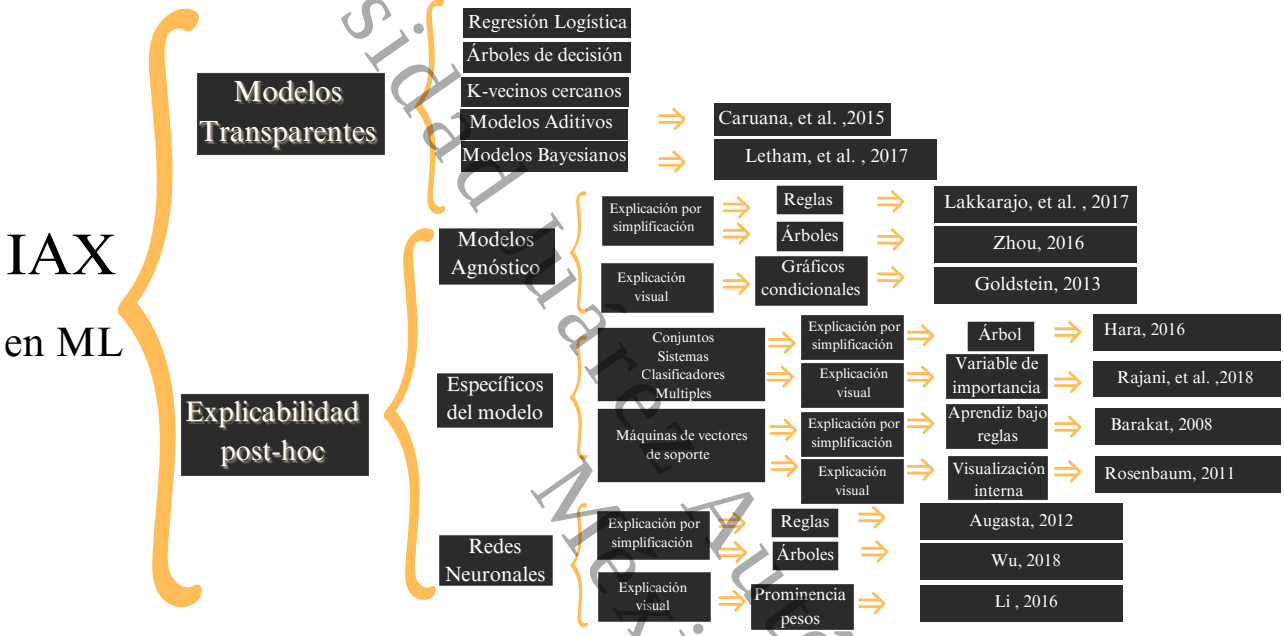

Figura 1.2. Comparación de algoritmos de explicativos contra precisión (Duval, 2019).

La interpretabilidad en modelos de aprendizaje automático ha cobrado relevancia en dominios críticos como el sector salud, donde no solo se requiere precisión predictiva, sino también comprensión y transparencia en la toma de decisiones. En este contexto, los enfoques de explicabilidad se pueden clasificar en dos grandes categorías: modelos intrínsecamente interpretables y métodos de explicabilidad*post-hoc*.

Por un lado, los modelos transparentes o intrínsecamente interpretables están diseñados para ser comprensibles desde su estructura. Entre estos, destacan los modelos aditivos generalizados aplicados en el ámbito médico por Caruana et al., 2015, los cuales permiten analizar la contribución individual de cada variable en la predicción. De manera similar, Letham et al., 2015 proponen clasificadores basados en reglas con fundamentos bayesianos, proporcionando interpretaciones explícitas mediante reglas fácilmente entendibles.

Por otro lado, los métodos de explicabilidad *post-hoc* buscan interpretar modelos complejos

sin modificar su estructura interna. Estos métodos pueden ser agnósticos al modelo o específicos de un tipo de modelo. En la categoría de métodos agnósticos, Goldstein et al., 2014 introducen los gráficos de efectos condicionales individuales (ICE), los cuales permiten visualizar el impacto de una variable sobre la predicción. Asimismo, Zhou y Hooker, 2016 proponen aproximar modelos complejos mediante árboles de decisión interpretables, facilitando una comprensión global del comportamiento del modelo.

En cuanto a los métodos específicos del modelo, diversas investigaciones se han enfocado en extraer conocimiento interpretable a partir de modelos complejos. Barakat y Diederich, 2008 desarrollan técnicas de extracción de reglas para máquinas de soporte vectorial, mientras que Rosenbaum et al., 2011 proponen visualizaciones mediante mapas de calor para interpretar modelos lineales. En el caso de redes neuronales, Augasta y Kathirvalavakumar, 2012 presentan métodos de extracción de reglas mediante ingeniería inversa, y Li et al., 2016 introducen técnicas de visualización para analizar representaciones internas en modelos de procesamiento de lenguaje natural.

Adicionalmente, existen enfoques híbridos que buscan combinar la capacidad predictiva de modelos complejos con interpretabilidad. M. Wu et al., 2018 proponen la regularización de modelos profundos mediante estructuras tipo árbol, promoviendo interpretabilidad sin sacrificar desempeño.

En este contexto, surgen herramientas modernas de explicabilidad ampliamente utilizadas como LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations), las cuales pertenecen a la categoría de métodos *post-hoc* agnósticos. LIME genera aproximaciones locales interpretables del modelo alrededor de una instancia específica, mientras que SHAP se basa en valores de Shapley provenientes de la teoría de juegos para asignar la contribución de cada variable a la predicción.

Por otra parte, los Explainable Boosting Machines (EBM) representan un enfoque intrínsecamente interpretable basado en modelos aditivos generalizados con interacciones, combinando interpretabilidad y alto desempeño predictivo. A diferencia de LIME y SHAP, que explican modelos ya entrenados, EBM permite comprender directamente el modelo sin necesidad de técnicas adicionales de interpretación.

En este trabajo se integran ambos enfoques: por un lado, se emplea EBM como modelo in-

terpretable, y por otro, se utilizan SHAP y LIME para analizar y explicar modelos más complejos, permitiendo una comparación entre interpretabilidad intrínseca y explicabilidad *post-hoc*. Esta combinación permite obtener una visión más completa del comportamiento de los modelos y fortalece la confianza en los resultados obtenidos.

Con base en la literatura revisada, los enfoques se agrupan en modelos intrínsecamente interpretables y métodos de explicabilidad *post-hoc*, como se muestra en la Tabla 1.2 y 1.3

Tabla 1.2. Modelos intrínsecamente interpretables.

Autor	Título	Resumen
Caruana et al., 2015	Intelligible Models for HealthCare	Modelos aditivos interpretables aplicados al sector salud.
Letham et al., 2015	Interpretable Classifiers Using Rules	Clasificadores basados en reglas bayesianas interpretables.
M. Wu et al., 2018	Beyond Sparsity	Enfoque híbrido que induce interpretabilidad en modelos profundos.

Tabla 1.3. Métodos de explicabilidad *post-hoc*.

Autor	Título	Tipo	Resumen
Goldstein et al., 2014	Peeking Inside the Black Box	Agnóstico	Visualización ICE del efecto de variables.
Zhou y Hooker, 2016	Single Tree Approximation	Agnóstico	Aproximación con árboles interpretables.
Hara y Hayashi, 2016	Tree Ensembles Interpretable	Específico	Simplificación de ensembles de árboles.
Barakat y Diederich, 2008	Rule Extraction SVM	Específico	Extracción de reglas desde SVM.
Rosenbaum et al., 2011	Heat Map SVM	Específico	Visualización mediante mapas de calor.
Augasta y Kathirvalavakumar, 2012	Rule Extraction NN	Específico	Extracción de reglas desde redes neuronales.
Li et al., 2016	Visualizing Neural Models	Específico	Visualización de representaciones internas en NLP.
Rajani y Mooney, 2018	Visual Explanations	Agnóstico	Ensamble de explicaciones visuales.

A pesar de los avances en interpretabilidad, aún existe un compromiso entre precisión y explicabilidad, lo que motiva la exploración de enfoques híbridos que permitan equilibrar ambos aspectos, particularmente en aplicaciones médicas donde la interpretabilidad es crítica.

1.4. Planteamiento del problema

En los últimos años, la adopción de técnicas de Aprendizaje automático ha experimentado un crecimiento exponencial en el ámbito de la salud, consolidándose como una herramienta crucial para la identificación de patrones complejos en grandes volúmenes de datos clínicos. A pesar de su elevado rendimiento predictivo, la mayoría de los algoritmos de vanguardia tales como las arquitecturas de aprendizaje profundo o los métodos de ensamble operan bajo la premisa de "caja negra". Esta opacidad estructural impide comprender el proceso lógico interno mediante el cual se transforman las variables de entrada en un diagnóstico o predicción específica, generando una barrera significativa para su validación metodológica.

Esta limitación representa un desafío crítico en el entorno médico, donde la toma de decisiones clínicas no solo exige una alta precisión, sino también un marco ético de transparencia, explicabilidad y causalidad. En patologías crónicas de alta incidencia y morbimortalidad, como la Enfermedad Renal Crónica (ERC) y la nefropatía diabética, la detección oportuna es el factor determinante para ralentizar el deterioro funcional y evitar la insuficiencia renal terminal. Si bien los modelos predictivos ofrecen un potencial sin precedentes para la identificación temprana de factores de riesgo, su falta de interpretabilidad provoca desconfianza en el personal médico, limitando su adopción en la práctica clínica real ante el temor de sesgos inadvertidos o correlaciones espurias en el entrenamiento del algoritmo.

Este escenario se agrava en contextos demográficos específicos, como en la población del sureste de México, donde la prevalencia de comorbilidades metabólicas presenta particularidades genéticas y socioculturales que exigen modelos adaptados a la realidad epidemiológica local. Por lo tanto, no basta con diseñar modelos altamente precisos sobre conjuntos de datos estandarizados internacionalmente; es imperativo descifrar cómo influyen las variables clínicas locales en cada predicción. De este modo surge la necesidad de implementar la Inteligencia Artificial Explicable (XAI), un paradigma emergente orientado a dotar de transparencia a los procesos de inferencia computacional.

Bajo esta premisa, la presente investigación aborda la brecha existente entre la precisión algorítmica y la utilidad médica mediante el desarrollo y análisis de modelos intrínsecamente interpretables (EBM) y la aplicación de metodologías de evaluación *post-hoc* (SHAP y LIME).

A través de este enfoque híbrido de explicabilidad, se busca transformar las cajas negras en sistemas de apoyo a la decisión médica que no solo optimicen la detección de nefropatías, sino que también revelen los factores clínicos subyacentes asociados a la progresión del daño renal, garantizando una herramienta auditable, segura y confiable para el sector salud.

1.5. Preguntas de investigación

1. ¿Cómo se pueden caracterizar y ponderar los factores de riesgo asociados al desarrollo de nefropatías en una población clínica mediante el uso de algoritmos de aprendizaje automático supervisado?
2. ¿Cuáles son las variables clínicas y biomarcadores que ejercen la mayor influencia y contribución marginal en las predicciones macroscópicas y locales de los modelos de clasificación implementados?
3. ¿En qué medida la convergencia de enfoques de explicabilidad intrínseca (EBM) y métodos *post-hoc* (SHAP y LIME) reduce la opacidad en los resultados y proporciona interpretaciones clínicamente descriptivos para el personal médico?
4. ¿Bajo qué condiciones los modelos intrínsecamente interpretables logran mantener un rendimiento predictivo estadísticamente equivalente al de arquitecturas complejas de “caja negra” en el análisis de datos clínicos?

1.6. Hipótesis

En la presente tesis se plantean las siguientes hipótesis:

“*Un modelo de aprendizaje automático intrínsecamente interpretable (EBM) permite alcanzar un rendimiento predictivo mayor al 80 % de Accuraccy en la detección de nefropatía en un dataset público, demostrando un comportamiento estadísticamente no inferior al de los modelos tradicionales de caja negra.*”

"*La aplicación de métodos post-hoc, como SHAP y LIME, sobre modelos de caja negra permite evaluar de forma estadística la contribución y el impacto individual de las variables clínicas asociadas al desarrollo de la nefropatía.*"

1.6.1. Objetivo general

Crear modelos predictivos y explicativos para nefropatía orientados al análisis de riesgo y la identificación de factores clínicos en diversas poblaciones, mediante el uso de algoritmos de Aprendizaje automático interpretable y técnicas de explicabilidad.

1.6.2. Objetivos específicos

1. Crear modelos algoritmos de clasificación tradicionales considerados de "caja negra" (Random forest y SVM) para establecer la línea base de rendimiento predictivo en la detección de nefropatías.
2. Crear un modelo basado en EBM para la predicción transparente y la evaluación tanto global como local de los factores de riesgo renal.
3. Aplicar los marcos de explicabilidad *post-hoc* SHAP y LIME para extraer los vectores de importancia y las contribuciones de cada variable a nivel local y global.
4. Evaluar cuantitativamente el contraste (*trade-off*) entre la precisión predictiva y la capacidad interpretativa de los enfoques implementados, contrastando las explicaciones algorítmicas con la literatura médica.

1.7. Alcances y limitaciones

El presente trabajo se enfocó en el desarrollo y análisis de modelos de aprendizaje automático interpretables aplicados a 2 conjuntos de datos clínicos: *(i)* pacientes con enfermedad renal crónica obtenidos del Chronic Kidney Disease dataset y *(ii)* pacientes con complicaciones renales por diabetes tipo 2 obtenidos del dataset DiabetIA. El estudio se centró unicamente en la identificación de factores asociados a nefropatías a partir de variables clínicas, signos vitales y resultados

de laboratorio disponibles en los conjuntos de datos utilizados.

Los resultados obtenidos dependieron de la calidad, tamaño y características de los conjuntos de datos analizados, por lo que las conclusiones derivadas del estudio deben interpretarse dentro del contexto de los datasets empleados. Asimismo, el estudio se limitó a la evaluación de determinados modelos de aprendizaje automático y técnicas de interpretabilidad, por lo que el uso de otros algoritmos o técnicas podrían arrojar resultados diferentes.

1.8. Justificación

La integración de sistemas basados en Aprendizaje automático en el área médica implica un cambio de paradigma en el diagnóstico asistido por computadora. Sin embargo, para que los médicos adopten estas tecnologías en escenarios reales, es indispensable que los algoritmos puedan explicar y justificar sus predicciones. En el ámbito de la nefrología, donde afecciones como la enfermedad renal crónica (ERC) presentan una evolución inicialmente asintomática y multifactorial, la detección oportuna de factores de riesgo es determinante para la supervivencia del paciente. En este escenario, un modelo predictivo que opere como “caja negra” resulta insuficiente, ya que los médicos requieren comprender la causalidad y el peso de los biomarcadores antes de prescribir intervenciones terapéuticas complejas o invasivas.

Por lo tanto, esta investigación aborda directamente la brecha entre la opacidad algorítmica y la necesidad de interpretabilidad clínica. Al proponer un enfoque híbrido que incluye modelos transparentes por diseño (EBM) con técnicas de explicabilidad *post-hoc* (SHAP y LIME), este trabajo busca alcanzar precisión predictiva y también proveer explicaciones tanto locales como globales que sean matemáticamente consistentes y clínicamente interpretables.

Esto permitirá identificar de forma transparente perfiles de riesgo específicos, y apoyar al personal de salud como soporte a la decisión clínica.

1.9. Metodología

Para el cumplimiento de los objetivos planteados y la validación de las hipótesis, la investigación se estructuró en las siguientes etapas metodológicas:

Fase 1: Revisión sistemática de la literatura. Consiste en la fundamentación del estado del arte, identificando los biomarcadores clínicos más reportados en la literatura nefrológica y analizando las tendencias vigentes en la aplicación de marcos de XAI en el área de la salud.

Fase 2: Preprocesamiento de los conjuntos de datos clínicos. Implica el análisis exploratorio de ambos datasets de riesgo renal, la gestión de valores faltantes mediante técnicas de imputación, la detección de valores atípicos y la codificación de variables categóricas, garantizando la calidad de los datos de entrada.

Fase 3: Diseño e implementación de modelos de caja negra. Desarrollo y ajuste de hiperparámetros de los modelos base "opacos" (Random forest y SVM), estableciendo el límite superior de rendimiento predictivo (*baseline*) mediante validación cruzada.

Fase 4: Diseño del modelo intrínsecamente interpretable. Entrenamiento y parametrización de EBM, modelando de forma aislada el impacto local y global de las variables para obtener transparencia nativa desde el propio algoritmo.

Fase 5: Configuración de las técnicas *post-hoc* SHAP y LIME sobre los modelos de caja negra para extraer las atribuciones locales y globales de las características.

Fase 6: Finalmente, se realizó un análisis comparativo y de convergencia entre las explicaciones generadas por EBM, SHAP y LIME, evaluando su consistencia mutua y su coherencia con las guías de práctica clínica renal.

1.10. Contribución y resultados esperados

En el área de las Ciencias de la Computación, la principal contribución de este trabajo radica en el diseño y validación de modelos interpretables y la evaluación *post-hoc* agnóstica de dichos modelos. A diferencia de los enfoques tradicionales que se limitan al uso de una única herramienta de explicabilidad, esta investigación aporta un esquema de validación entre modelos que permite contrastar la consistencia matemática de las explicaciones locales y globales generadas por EBM, SHAP y LIME sobre dos conjuntos de datos diferentes. Esto robustece la confiabilidad

de los vectores de importancia de características, sentando un precedente metodológico para la auditoría de modelos predictivos complejos.

En el ámbito de la medicina y el sector salud, el resultado esperado es el desarrollo de un modelo predictivo-explicativo con alta sensibilidad para la detección oportuna de factores de riesgo asociados a la enfermedad renal crónica y la nefropatía diabética. Más allá de la precisión cuantitativa, el modelo proporciona perfiles explicativos acerca de las particularidades clínicas de los conjuntos de datos utilizados. Esto constituye un sistema de soporte a la decisión médica auditable, facilitando al personal de salud la comprensión de las patologías y promoviendo el diseño de estrategias de intervención temprana personalizadas.

Finalmente, como evidencia de la validez de los hallazgos derivados de esta investigación, se realiza la transferencia de conocimiento a la comunidad académica internacional mediante la publicación de los resultados intermedios y finales en revistas científicas indexadas especializadas.

1.11. Cronograma

Tabla 1.4. Cronograma de actividades del proyecto.

No.	Actividad	S1	S2	S3	S4	S5	S6
1	Revisión de literatura	✓	✓	✓	✓	✓	✓
2	Redacción del protocolo	✓	✓				
3	Recopilación y compresión de los datos		✓	✓	✓		
4	Preparación de los datos		✓	✓	✓		
5	Análisis de algoritmos de caja negra		✓	✓	✓		
6	Análisis de algoritmos interpretables		✓	✓	✓	✓	
7	Modelado			✓	✓	✓	✓
8	Evaluación				✓	✓	✓
9	Redacción y envío de artículo			✓	✓	✓	✓
10	Redacción y revisión de la tesis			✓	✓	✓	✓
11	Presentación y defensa de la tesis						✓

Referencias del capítulo

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters*, *35*(2), 131-150. https://doi.org/10.1007/s11063-011-9208-y

Barakat, N., & Diederich, J. (2008). Eclectic Rule-Extraction from Support Vector Machines. *International Journal of Computer and Information Engineering*, *2*(5), 1672-1675.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth; Brooks/Cole.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2788613

Duval, A. (2019). *Explainable Artificial Intelligence (XAI)* (Scholarly Report). Mathematics Institute, The University of Warwick.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2014). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1). https://doi.org/10.1080/10618600.2014.907095

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)* (inf. téc.). Defense Advanced Research Projects Agency (DARPA).

Hara, S., & Hayashi, K. (2016). Making Tree Ensembles Interpretable.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second). Springer Science & Business Media.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350-1371. https://doi.org/10.1214/15-AOAS848

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and Understanding Neural Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 188-197.

Lundberg, S. M., & Lee, S.-I. (2017b). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 4765-4774. https://arxiv.org/abs/1705.07874

Mitchell, T. M. (2019). *Machine Learning* (2.ª ed.). McGraw-Hill.

Rajani, N. F., & Mooney, R. J. (2018). Ensembling Visual Explanations. En *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 155-172). Springer. https://doi.org/10.1007/978-3-319-98131-4_7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You¿': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rosenbaum, L., Hinselmann, G., & Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. *Journal of Cheminformatics*, *3*, 11. https://doi.org/10.1186/1758-2946-3-11

Russell, S., & Norvig, P. (2022). *Inteligencia artificial: Un enfoque moderno* (4.ª ed.). Pearson Educación.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second). MIT Press.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond Sparsity: Tree Regularization of Deep Models for Interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 1670-1678.

Zhou, Y., & Hooker, G. (2016). Interpreting Models via Single Tree Approximation.

Capítulo 2

Análisis estadístico de variables relacionadas con insuficiencia renal

Análisis estadístico de variables relacionadas con insuficiencia renal

Simón Javier Hernández-Gaspar; Oscar Chávez-Bosquez; Betania Hernández-Ocaña

Resumen La insuficiencia renal crónica (ERC, por Enfermedad Renal Crónica) es una enfermedad devastadora que no tiene cura, que requiere de un trasplante de riñón y que, si no es tratada adecuadamente, puede llevar a la muerte al paciente. Para realizar un estudio exploratorio de la variables relacionadas con esta enfermedad, utilizó un dataset público que fue obtenido del repositorio UCI (UC Machine Learning Repository), el cual consta de 400 instancias, 24 atributos y una clase con dos salidas, enfermo o no enfermo. El análisis se desarrolló en la plataforma Kaggle, la cual cuenta con una máquina virtual donde es posible ejecutar notebooks Jupyter escritas en el lenguaje de programación Python. En particular se utilizó la biblioteca Pandas para manipular y analizar los datos, así como las bibliotecas Matplotlib y Seaborn para visualizar cada una de las variables por medio de gráficos. En este caso se analizan a detalle cada una de las 11 variables numéricas incluidas en el dataset, y en particular, dos variables nominales: diabetes mellitus (`dm`) y la clase `class`. La variable `dm` resultó tener particularidades en sus valores. Se observa que todas las variables tienen valores nulos y valores faltantes indicados con un signo de interrogación (`?`), exceptuando la clase. Se concluye que el dataset está desbalanceado (250 enfermos contra 150 no enfermos) y que requiere imputación de datos en todas las variables antes de aplicar algoritmos de aprendizaje automático para desarrollar modelos de pronóstico de la enfermedad.

Palabras clave: Análisis exploratorio de datos, dataset, insuficiencia renal, plataforma Kaggle.

Institución de adscripción: División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México.

2.1. Introducción

La Enfermedad Renal Crónica (ERC) se define como la pérdida gradual y progresiva de la función renal, caracterizada por la incapacidad de los riñones para filtrar adecuadamente las toxinas y el exceso de líquidos de la sangre. Esta afección representa un problema de salud pública global debido a su naturaleza asintomática en las etapas iniciales, lo que retrasa su diagnóstico oportuno y eleva los costos de atención médica al evolucionar hacia fases terminales (Arias Rodríguez et al., 2013).

Para que un paciente desarrolle insuficiencia renal crónica, la patología progresa a través de cinco estadios evolutivos, los cuales transitan desde un daño renal leve con filtración glomerular conservada hasta el quinto estadio (Arias Rodríguez et al., 2013). En esta última etapa, el deterioro es severo y compatible con la pérdida de la función de los órganos, por lo que el paciente requiere de forma obligatoria ingresar a una terapia de reemplazo renal, tales como la diálisis peritoneal domiciliaria, la hemodiálisis o el trasplante renal (Levey et al., 2007).

A nivel local, la diabetes mellitus constituye la principal causa de ERC en el estado de Tabasco; sin embargo, existen otros factores de riesgo determinantes en la región que contribuyen a su prevalencia, tales como la hipertensión arterial, infecciones urinarias crónicas, enfermedad poliquística renal, litiasis renal, trastornos hipertensivos del embarazo (preeclampsia) y lupus eritematoso sistémico (Organización Panamericana de la Salud, 2026).

El impacto epidemiológico en la entidad es evidente: durante el periodo comprendido entre los años 2018 y 2021, en Tabasco se registraron oficialmente 671 casos de insuficiencia renal. Cabe destacar que los registros se duplicaron en el año 2020 y se triplicaron durante el 2021 en comparación con el comportamiento observado en 2018 y 2019. Asimismo, la mortalidad asociada a complicaciones renales superó las mil defunciones acumuladas en dicho periodo de cuatro años (Instituto Mexicano del Seguro Social, 2013).

Por lo tanto, resulta de gran interés comprender el comportamiento epidemiológico y clínico de la ERC, con el objetivo de diseñar estrategias de prevención y ofrecer un tratamiento oportuno a los ciudadanos que presentan una alta probabilidad de desarrollar esta patología.

Para este análisis se utilizó un conjunto de datos público obtenido del repositorio UCI (*UC Machine Learning Repository*), el cual consta de 400 registros de pacientes y 25 atributos en

total: 24 variables predictoras (11 numéricas y 13 nominales) y una variable de clase con dos posibles salidas (enfermo o no enfermo).

El desarrollo experimental se llevó a cabo en la plataforma Kaggle, la cual provee un entorno virtual optimizado para la ejecución de *notebooks* de Jupyter basadas en el lenguaje de programación Python (Géron, 2022). En particular, se empleó la biblioteca Pandas para la manipulación y el análisis estructurado de los datos, así como las herramientas Matplotlib y Seaborn para la exploración visual de las variables mediante gráficos estadísticos (Deitel y Deitel, 2019). Este enfoque metodológico permitió evaluar la correlación entre las características clínicas y determinar cuáles de ellas presentan un mayor impacto o poder discriminante para predecir la presencia o ausencia de la enfermedad.

2.2. Desarrollo

2.2.1. Objetivos

Objetivo General:

Realizar un análisis exploratorio del dataset Chronic Kidney Disease.

Objetivos específicos:

Implementar técnicas de estadística descriptiva para identificar el comportamiento de cada variable.

Crear gráficos para la visualización de la distribución de cada una de las variables.

2.3. Objeto de estudio

Para llevar acabo este estudio se utilizó el dataset Chronic Kidney Disease (Insuficiencia renal crónica), el cual consta de 400 instancias, 24 atributos y una clase con dos salidas, insuficiencia renal o no insuficiencia renal. De las 24 variables, 11 son numéricas y 14 son nominales. Al momento de descargar el dataset no tiene el formato csv, es por ello por lo que, al convertirlo, las variables se cargan automáticamente en formato nominal. ese fue uno de los primeros problemas en el dataset, Para solucionar dichos problemas en las columnas numéricas se utilizó la

Tabla 2.1. Descripción de las variables del dataset Chronic Kidney Disease

No.	Variable	Significado	Valores	Unidad/Rango	Faltantes
1	age	Edad	Años	Numérica	9
2	bp	Presión arterial	mm/Hg	Numérica	12
3	sg	Gravedad específica	1.005 – 1.025	Nominal	47
4	al	Albumina	0, 1, 2, 3, 4, 5	Nominal	46
5	su	Azúcar	0, 1, 2, 3, 4, 5	Nominal	49
6	rbc	Glóbulos rojos	normal, anormal	Nominal	152
7	pc	Células de pus	normal, anormal	Nominal	65
8	pcc	Grupos de células pus	presente, no presente	Nominal	4
9	ba	Bacterias	presente, no presente	Nominal	4
10	bgr	Glucosa en sangre	mg/dL	Numérica	44
11	bu	Urea en la sangre	mg/dL	Numérica	19
12	sc	Suero de creatina	mg/dL	Numérica	17
13	sod	Sodio	mEq/L	Numérica	87
14	pot	Potasio	mEq/L	Numérica	88
15	hemo	Hemoglobina	gms	Numérica	52
16	pcv	Volumen glóbulos rojos	Pendiente	Numérica	71
17	wbcc	Núm. glóbulos blancos	Cell/cmm	Numérica	106
18	rbcc	Núm. glóbulos rojos	millones/cmm	Numérica	131
19	htn	Hipertensión	sí, no	Nominal	2
20	dm	Diabetes mellitus	sí, no	Nominal	2
21	cad	Arteriopatía coronaria	sí, no	Nominal	2
22	appet	Apetito	bueno, malo	Nominal	1
23	pe	Edema de piel	sí, no	Nominal	1
24	ane	Anemia	sí, no	Nominal	1
25	class	Clase (Target)	ckd, notckd	Nominal	0

instrucción `astype('float64')` de la biblioteca Pandas. En la Tabla 4.1 se muestra cada una de las variables, con su valor de medida y valores faltantes.

2.4. Metodología

Se usó el análisis exploratorio (EDA, por las siglas en Inglés de *Exploratory Data Analysis*), el cual consiste en los siguientes pasos:

- Identificación de variables.
- Análisis univariado (estadística descriptiva y visualización).
- Análisis Bivariado.

- Valores faltantes.
- Valores atípicos (*outliers*).

En este trabajo se realizaron específicamente las 2 primeras fases del EDA: identificación de variables y el análisis univariado (Mendenhall et al., 2014).

Identificación de variables: Este paso consiste en poder determinar la variable de entrada junto con la variable de salida, ya que es muy importante la calidad del valor de entrada porque eso dará una buena salida. Después hay que saber el tipo de dato y la categoría de cada variable (Myers et al., 2012).

Análisis univariado: En este paso se realiza la exploración de variable por variable, y este tipo de exploración depende de qué tipo de variable se trate (Myers et al., 2012):

- Variables numéricas: este tipo de variables se puede analizar usando estadística descriptiva.
- Variables nominales: en este tipo de variables se usan gráficas de barras y tablas de frecuencia para su análisis e interpretación; también es posible leer como porcentaje los valores en cada categoría.

Las técnicas de exploración de datos tienen como objetivo maximizar el conocimiento de un conjunto de datos, detectar valores atípicos, anomalías y probar suposiciones adyacentes. También pueden minimizar el tiempo de análisis del conjunto de datos, utilizando una rama de las Matemáticas como lo es la Estadística descriptiva. Asimismo, la visualización de las variables ayuda a interpretar de manera gráfica cada variable con mayor facilidad.

2.5. Fases del desarrollo

El análisis univariado de las variables numéricas se realiza con ayuda de la estadística descriptiva. A continuación se describen las variables cuantitativas del dataset.

Análisis de la variable age

La edad de las instancias del dataset tiene un valor mínimo de 2 años y un valor máximo de 90 años. En la Figura 2.1a se observa que la mayor concentración de los pacientes están entre

50 y 70 años. En la Figura 2.1b se muestran valores atípicos y la media color verde. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   51.48
Mediana: 55.0
Moda: 60
Name: age, dtype: float64
Desviación estándar: 17.17
```

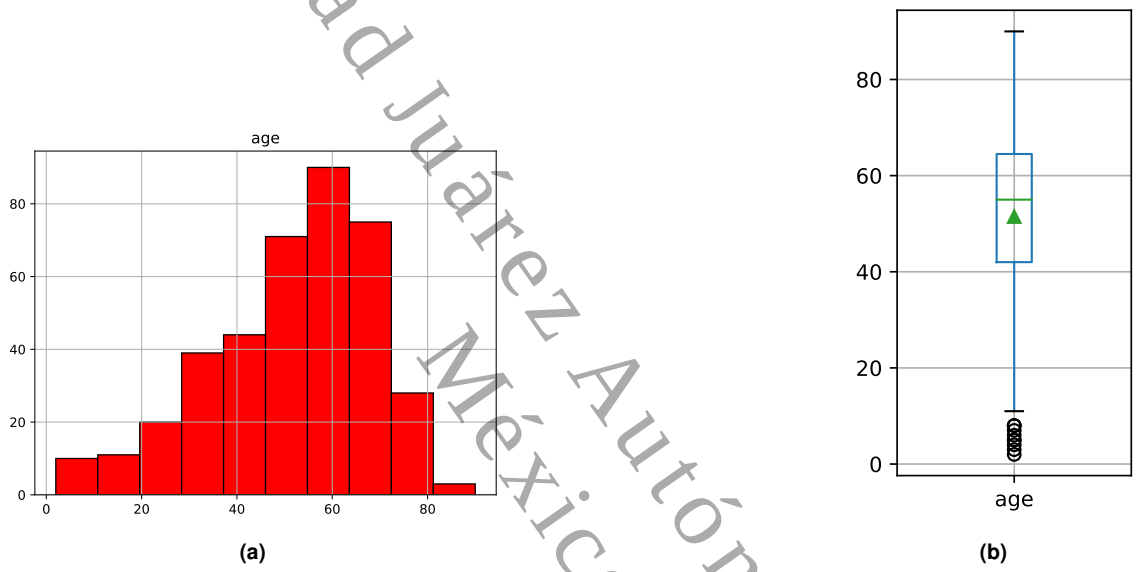

Figura 2.1. (a) Histograma y (b) diagrama de cajas de la variable `age`.

Análisis de la variable `bp`

La presión arterial (`bp` por las siglas en Inglés de *blood preasure*) es la fuerza que ejerce contra la pared arterial la sangre que circula por las arterias. La presión arterial incluye dos mediciones: la presión sistólica, que se mide durante el latido del corazón (momento de presión máxima), y la presión diastólica, que se mide durante el descanso entre dos latidos (momento de presión mínima). Primero se registra la presión sistólica y luego la presión diastólica, por ejemplo: 120/80. También se llama presión sanguínea arterial y tensión arterial. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   76.47
Mediana: 80.0
```

```
Moda: 80.0
Name: bp, dtype: float64
Desviación estandar: 13.68
```

En la Figura 2.2a se observa que la mayor concentración de los pacientes están entre 70 y 90. En la Figura 2.2b se muestran valores atípicos y la media en color verde.

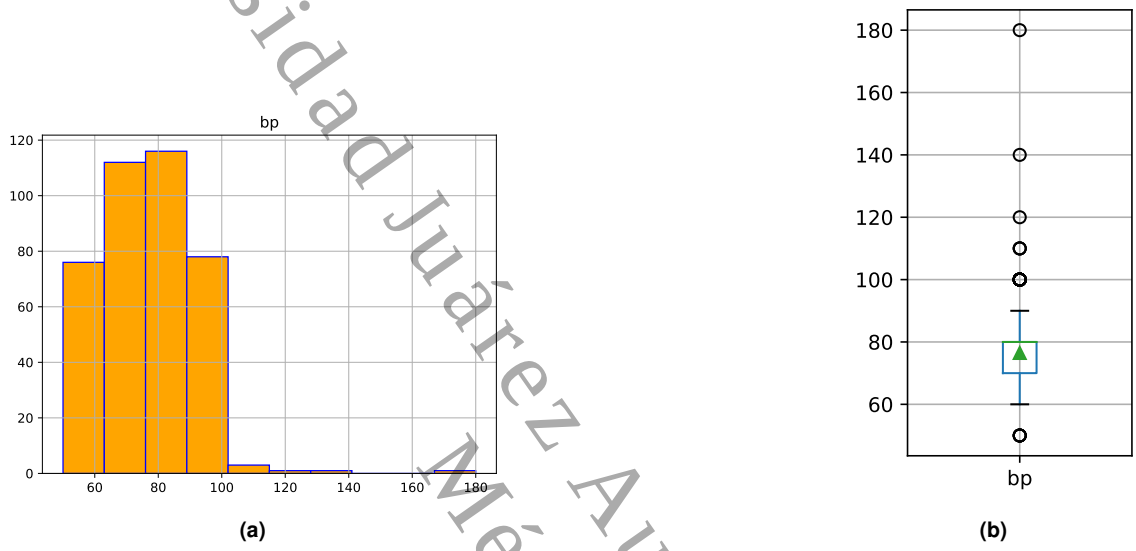

Figura 2.2. (a) Histograma y (b) diagrama de cajas de la variable `bp`.

Análisis de la variable `bgr`

La glucosa en sangre (*bgr* por las siglas en Inglés de *blood glucose random*) es el azúcar principal que se encuentra en la sangre. Es la principal fuente de energía del cuerpo y proviene de los alimentos que se consume. El cuerpo descompone la mayor parte de ese alimento en glucosa y la libera en el torrente sanguíneo. La Organización Mundial de la Salud (OMS) indica que un rango normal de azúcar en la sangre, en ayunas, se encuentra entre los 70 y 100 mg/dl (miligramos por decilitro); un rango de entre 101 y 125 mg/dl es una glucemia basal alterada, algo que se podría considerar como pre diabetes; y un rango de 126 mg/dl o más es diagnóstico de diabetes. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   148.04
Mediana: 121.0
Moda: 99.0
```

```
Name: bgr, dtype: float64
Desviación estandar: 79.28
```

En la Figura 2.3a se observa que la mayor concentración de los pacientes están entre 90 y 150. En la Figura 2.3b se muestran valores atípicos y la media color verde.

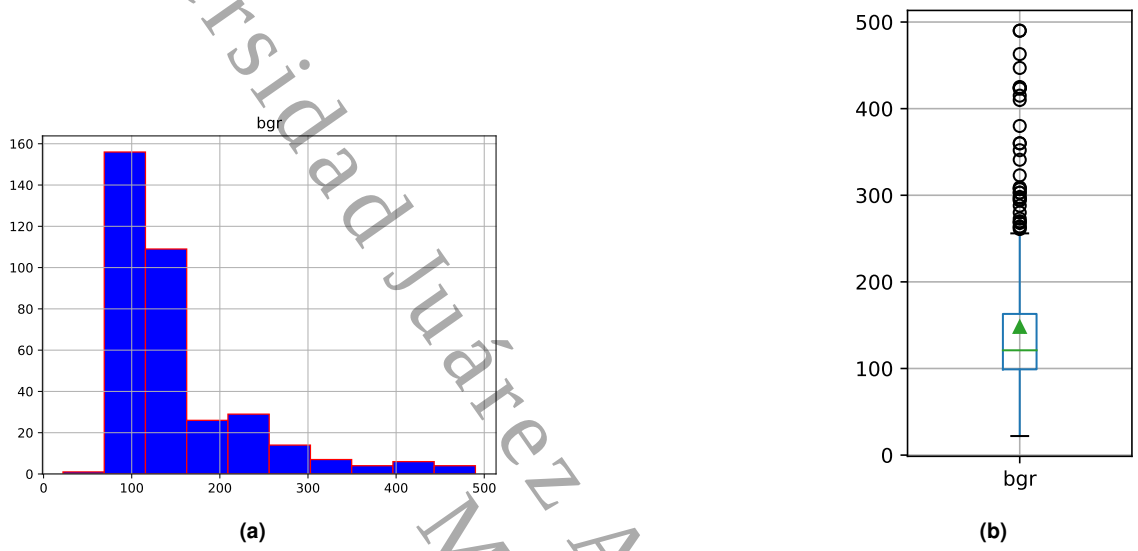

Figura 2.3. (a) Histograma y (b) diagrama de cajas de la variable `bgr`.

Análisis de la variable `bu`

El análisis de nitrógeno ureico en sangre (*bu*, por las siglas en Inglés de *blood urea*) mide la cantidad de nitrógeno en la sangre que proviene de un producto de desecho, llamado urea. La urea se produce cuando se descompone la proteína en el cuerpo. La urea se produce en el hígado y se excreta del cuerpo en la orina. Este estudio se hace cuando existe sospecha que de un problema renal o si particularmente se padece una afección crónica, como diabetes o presión arterial alta. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   57.43
Mediana: 42.0
Moda: 46.0
Name: bu, dtype: float64
Desviación estandar: 50.5
```

En la Figura 2.4a se observa que la mayor concentración de los pacientes están entre 0 y 50.

En la Figura 2.4b se muestran valores atípicos y la media color verde.

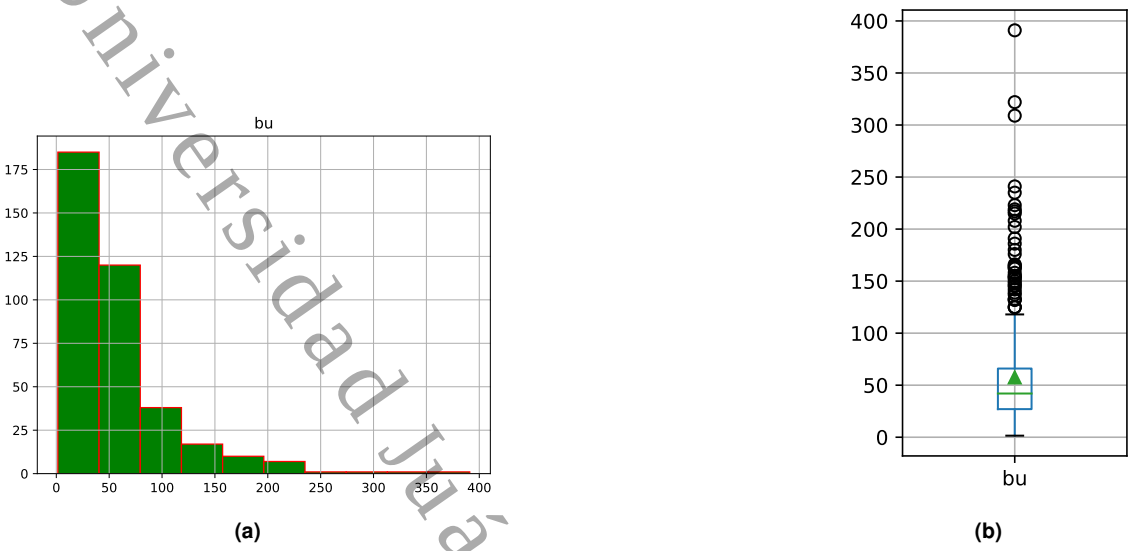

Figura 2.4. (a) Histograma y (b) diagrama de cajas de la variable `bu`.

Análisis de la variable `sc`

La creatinina (`sc` por las siglas en Inglés de *serum creatinine*) es una sustancia que se produce a partir de la creatina y el creatinfosfato como resultado de los procesos metabólicos musculares. Muestra el trabajo que realizan los riñones, encargados de absorber esta sustancia en su fase final y luego eliminarla por la orina. Un aumento o descontrol de la creatinina puede indicar problemas renales. Los valores normales varían entre 0,6 y 1,1 mg/dL en las mujeres y 0,7 y 1,3 mg/dL en los hombres. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   3.07
Mediana: 1.3
Moda: 1.2
Name: sc, dtype: float64
Desviación estandar: 5.74
```

En la Figura 2.5a se observa que la mayor concentración de los pacientes están entre 0 y 8. En la Figura 2.5b se muestran valores atípicos y la media color verde.

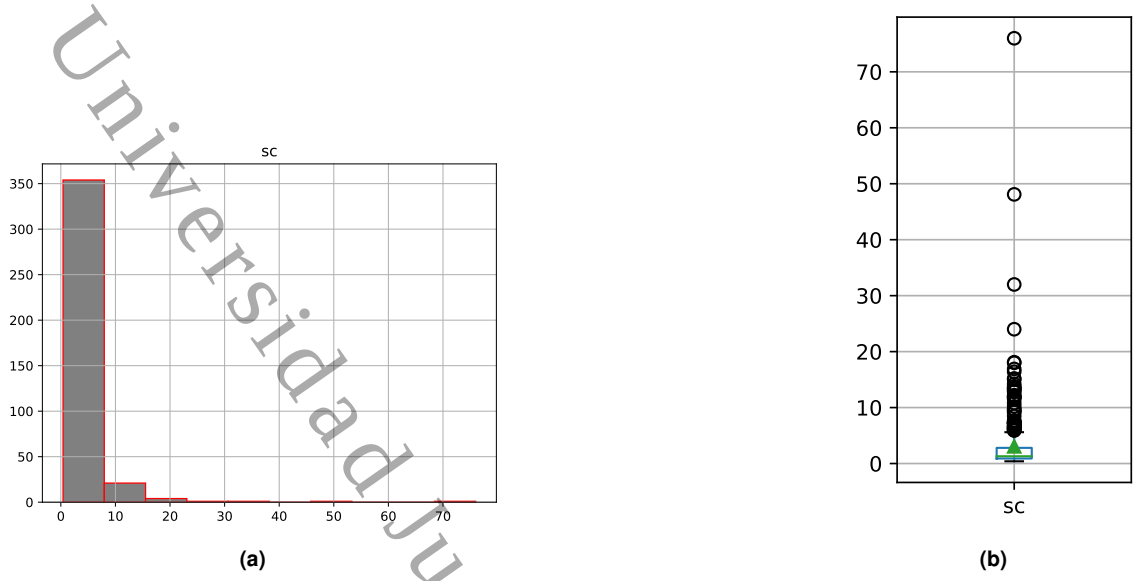

Figura 2.5. (a) Histograma y (b) diagrama de cajas de la variable `sc`.

Análisis de la variable `sod`

El sodio (`sod` por las siglas en Inglés de *sodium*) tiene una función clave en el cuerpo. Ayuda a mantener una presión arterial normal y apoya el trabajo de los nervios y músculos, a la vez que regula el equilibrio de líquidos en el cuerpo. Un nivel normal de sodio en la sangre oscila entre 135 y 145 miliequivalentes por litro (mEq/L). Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   137.53
Mediana: 138.0
Moda: 135.0
Name: sod, dtype: float64
Desviación estandar: 10.41
```

En la Figura 2.6a se observa que la mayor concentración de los pacientes están entre 130 y 145. En la Figura 2.6b se muestran valores atípicos y la media color verde.

Análisis de la variable `pot`

El potasio (`pot` por las siglas en Inglés de *potassium*) es un tipo de electrolito. Los electrolitos son minerales con carga eléctrica que ayudan a controlar los niveles de líquidos y el balance de ácidos y bases (equilibrio pH) en el cuerpo. También ayuda a controlar la actividad de los

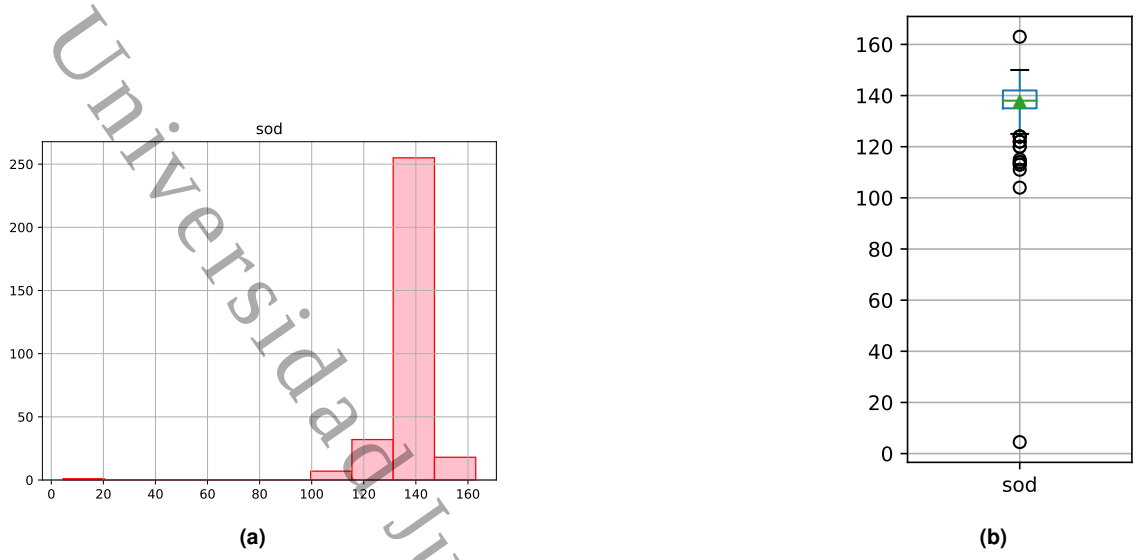

Figura 2.6. (a) Histograma y (b) diagrama de cajas de la variable `sod`.

músculos y los nervios, también cumplen otras funciones importantes. Los niveles de potasio con frecuencia cambian con los niveles de sodio. Cuando los niveles de sodio aumentan, los niveles de potasio disminuyen, y cuando los niveles de sodio disminuyen, los niveles de potasio aumentan. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   4.63
Mediana: 4.4
Moda: 3.5, 1, 5.0
Name: pot, dtype: float64
Desviación estandar: 3.19
```

En la Figura 2.7a se observa que la mayor concentración de los pacientes están entre 3 y 6. En la Figura 2.7b se muestran valores atípicos y la media color verde.

Análisis de la variable `hemo`

La hemoglobina (`hemo` por las siglas en Inglés de *hemoglobin*) es una proteína del interior de los glóbulos rojos que transporta oxígeno desde los pulmones a los tejidos y órganos del cuerpo; además, transporta el dióxido de carbono de vuelta a los pulmones. Por lo general, la prueba para medir la cantidad de hemoglobina en la sangre forma parte del recuento sanguíneo completo (RSC). Los valores obtenidos por el análisis exploratorio fueron:

Figura 2.7. (a) Histograma y (b) diagrama de cajas de la variable `pot`.

```
Media:   12.53
Mediana: 12.65
Moda: 15.0
Name: hemo, dtype: float64
Desviación estandar: 2.91
```

En la Figura 2.8a se observa que la mayor concentración de los pacientes están entre 13 y 16. En la Figura 2.8b se muestran valores atípicos y la media color verde.

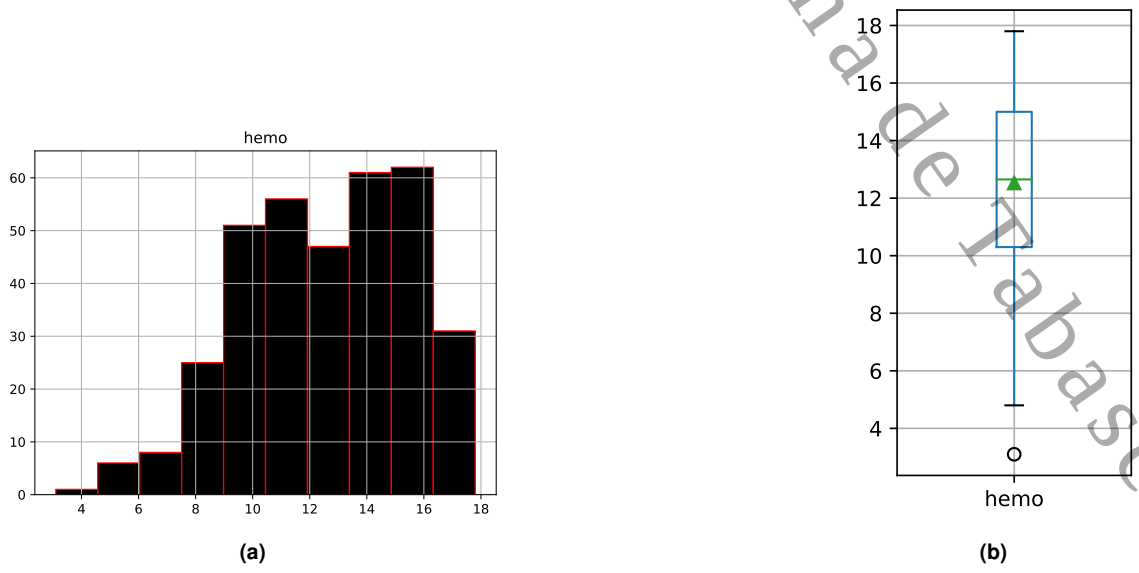

Figura 2.8. (a) Histograma y (b) diagrama de cajas de la variable `hemo`.

Análisis de la variable `pcv`

El volumen de glóbulos rojos (`pcv` por las siglas en Inglés de *packed cell volume*) se obtiene mediante un examen que ayuda a diagnosticar diferentes tipos de anemia (bajo número de glóbulos rojos) y otros problemas de salud que afectan los glóbulos rojos. Mide el número de glóbulos rojos que existen en una muestra de sangre. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   38.88
Mediana: 40.0
Moda: 41.0, 1, 52.0
Name: pcv, dtype: float64
Desviación estandar: 8.99
```

En la Figura 2.9a se observa que la mayor concentración de los pacientes están entre 35y 48. En la Figura 2.9b se muestran valores atípicos y la media color verde.

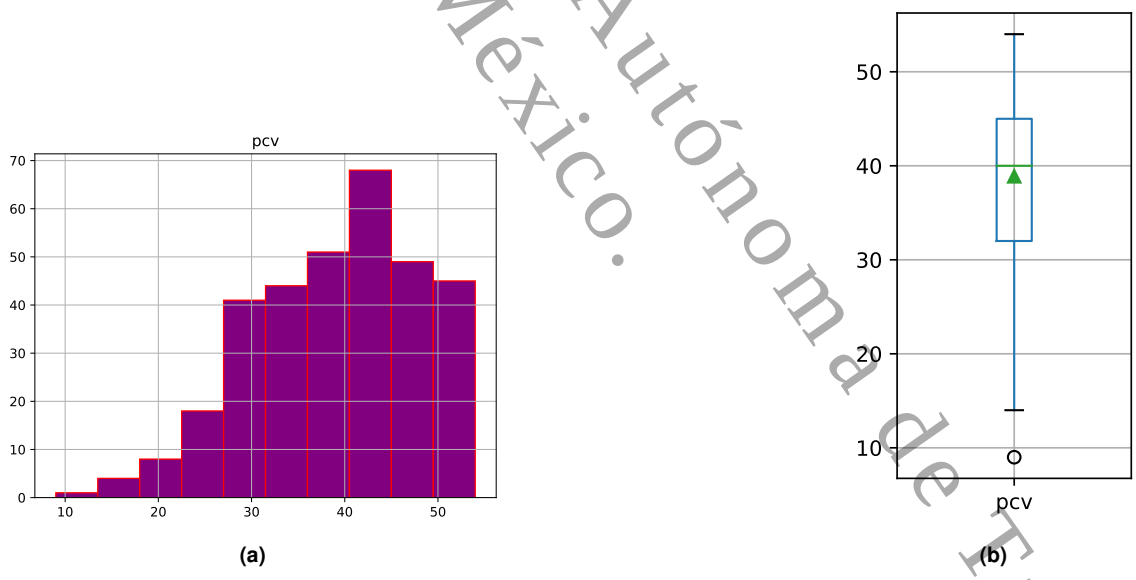

Figura 2.9. (a) Histograma y (b) diagrama de cajas de la variable `pcv`.

Análisis de la variable `wbcc`

Es un análisis para medir la cantidad de glóbulos blancos (`wbcc` por las siglas en Inglés de *white blood cell count*) que tiene en la sangre. Los glóbulos blancos también se llaman leucocitos.

La médula ósea produce glóbulos blancos y los libera al torrente sanguíneo. Estos glóbulos ayudan a combatir las infecciones. Forman parte del sistema inmunitario del cuerpo, que lo mantiene saludable y lo ayuda a mejorarse cuando se enferma. Los glóbulos blancos destruyen virus, hongos o bacterias extraños que ingresan al cuerpo. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   8406.12
Mediana: 8000.0
Moda: 9800.0
Name: wbcc, dtype: float64
Desviación estandar: 2944.47
```

En la Figura 2.10a se observa que la mayor concentración de los pacientes están entre 35y 48. En la Figura 2.10b se muestran valores atípicos y la media color verde.

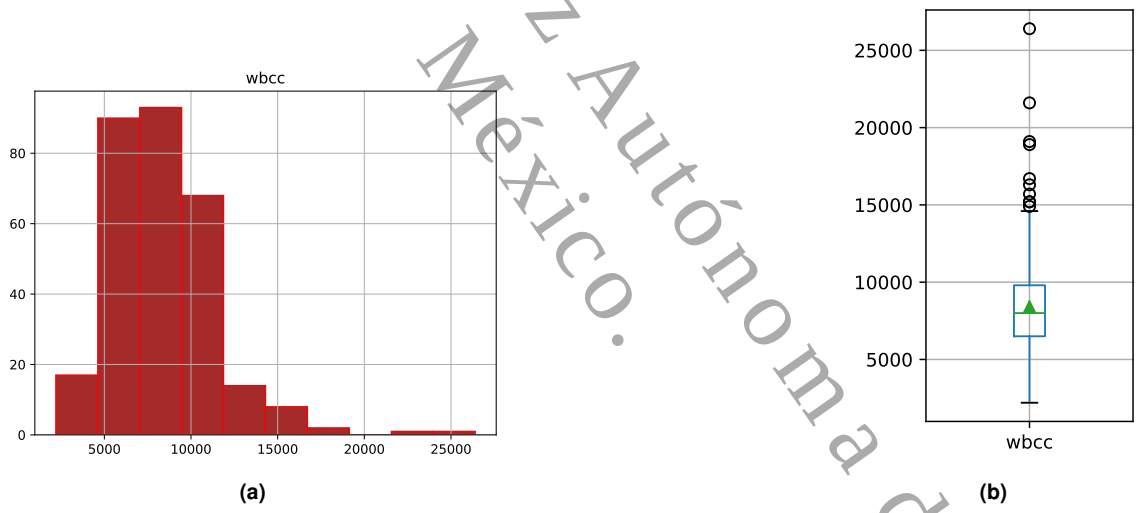

Figura 2.10. (a) Histograma y (b) diagrama de cajas de la variable `wbcc`.

Análisis de la variable `rbcc`

Es una prueba para medir la cantidad de glóbulos rojos (`rbcc` por las siglas en Inglés de *red blood cell count*), o eritrocitos, en la sangre. Los glóbulos rojos cumplen una función esencial en el transporte de oxígeno desde los pulmones hacia el resto del cuerpo, así como en llevar el dióxido de carbono de regreso a los pulmones para exhalarlo. Por lo general, el recuento de glóbulos rojos se hace como parte de un hemograma completo. Esta es una prueba de detección que permite

revisar diversas afecciones médicas. Los valores obtenidos por el análisis exploratorio fueron:

```
Media:   471
Mediana: 4.8
Moda: 5.2
Name: rbcc, dtype: float64
Desviación estandar: 1.03
```

En la Figura 2.11a se observa que la mayor concentración de los pacientes están entre 4.5 y 5.5. En la Figura 2.11b se muestran valores atípicos y la media color verde.

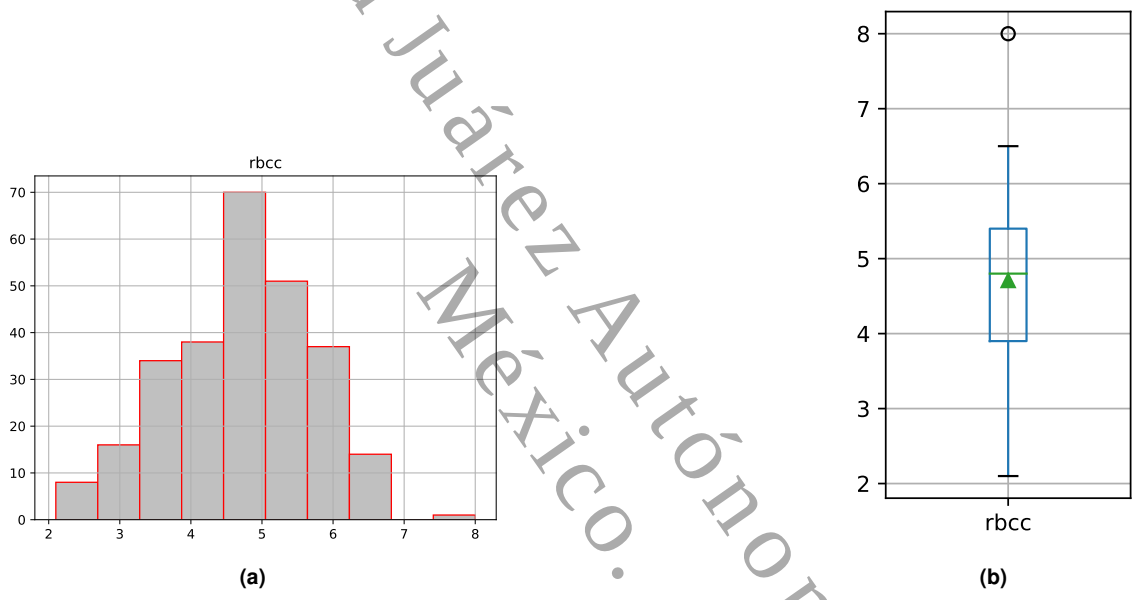

Figura 2.11. (a) Histograma y (b) diagrama de cajas de la variable `rbcc`.

Una vez analizadas las propiedades numéricas del conjunto de datos, se procede al estudio de las variables categóricas o cualitativas. Para este análisis, se emplean distribuciones de frecuencias y representaciones gráficas de barras, permitiendo identificar la prevalencia de los atributos clínicos y las características demográficas de los pacientes en el dataset.

Análisis de la variable `dm`

La diabetes mellitus se refiere a un grupo de enfermedades que afectan a la forma en que el cuerpo utiliza el azúcar en la sangre (glucosa). La glucosa es una fuente importante de energía para las células que constan los músculos y los tejidos. También es la principal fuente de com-

bustible del cerebro. La causa principal de la diabetes varía según el tipo. Pero no importa qué tipo de diabetes tengas, puede provocar un exceso de azúcar en la sangre. Demasiado azúcar en la sangre puede provocar graves problemas de salud, como dañar los vasos sanguíneos y los riñones. El análisis de esta variable categórica es:

```
dm
no    261
yes 136
NaN 2
yes 1
Name: count, dtype: int64
```

En la Figura 2.12 se observa que la cantidad de pacientes que tienen la enfermedad es mas grande de las que no la tienen, de igual manera podemos observar los valores nulos y la irregularidad. Ésta se refiere a que hay un “yes” mal escrito, tiene un espacio en blanco, por eso hay un valor adicional en la gráfica.

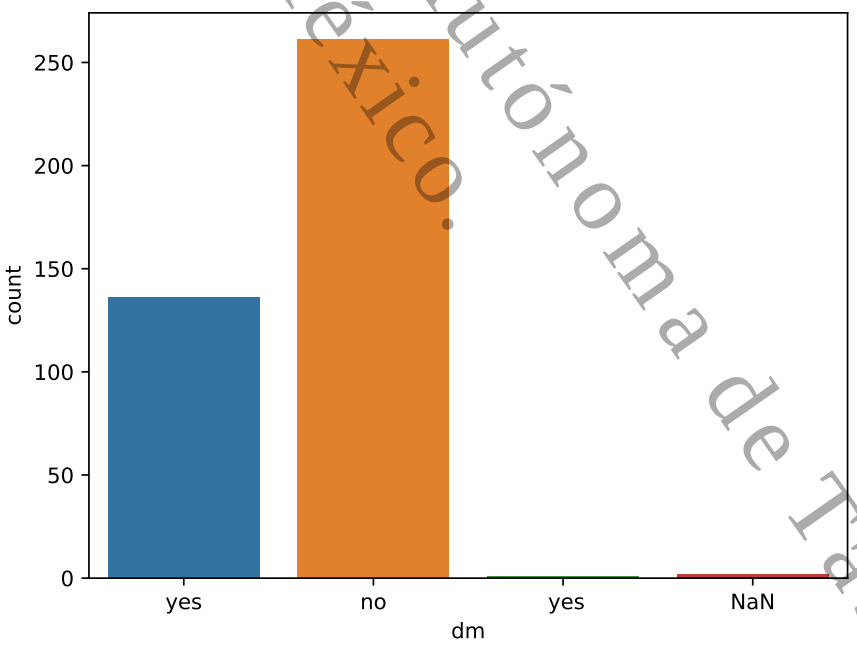

Figura 2.12. Distribución de frecuencias de la variable diabetes mellitus (`dm`).

Análisis de la variable `class`

La clase (*class*) representa uno de los dos diagnósticos que un paciente puede tener: insuficiencia renal o no insuficiencia renal (`ckd`, `notckd`).

```
class
ckd    250
notckd 150
Name: count, dtype: int64
```

En la Figura 2.13 se observa que el número de pacientes que tienen y las que no tienen la enfermedad.

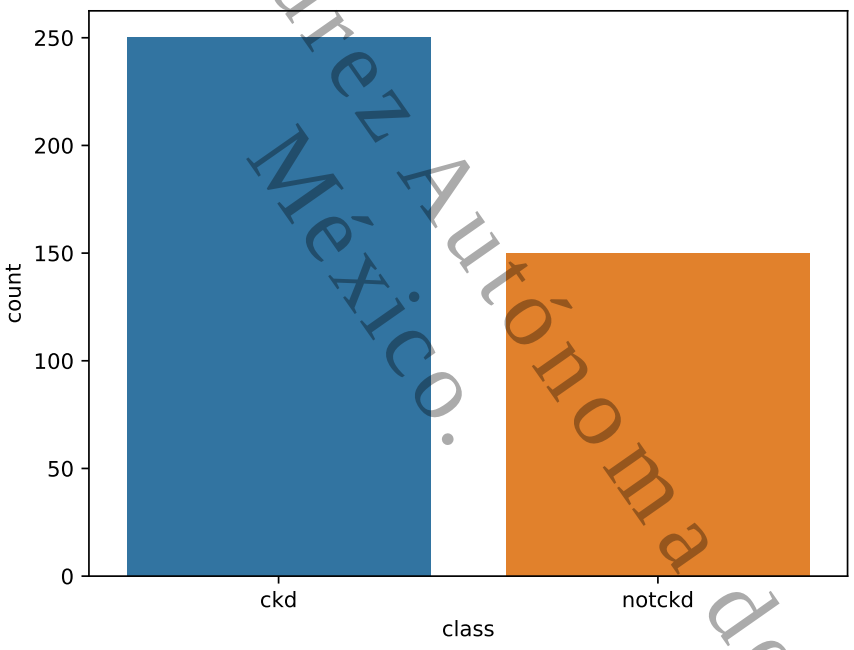

Figura 2.13. Distribución de frecuencias de la variable clase (`class`).

2.6. Resultados y discusión

Después de realizado el análisis univariado de las variables seleccionadas, se encontró que la variable `dm` es una variable nominal con dos posibles valores (*yes*, *no*). Sin embargo, en su estado original, se detectó que en algunas instancias el valor aparece como "`yes`" (con espacios

en blanco) o con valores nulos, lo que provoca que en su gráfico de barras se visualicen cuatro categorías en lugar de dos.

Para identificar estas inconsistencias, se utilizó el siguiente comando en Python:

```
dataset[(dataset['dm']!='yes') & (dataset['dm']!='no')]
```

Dicha instrucción permitió localizar las filas erróneas. Finalmente, mediante la instrucción `dataset.drop([30, 288, 297], inplace=True)`, se procedió a la eliminación de dichas instancias. En la Figura 2.14 se observa la distribución final de los pacientes con y sin la enfermedad.

Otro detalle encontrado fue en la variable `wc`, la cual aparece dentro el dataset como `wc` pero en la descripción se le cambió el nombre a `wbcc`. Realmente, buscando las siglas en Inglés se encontró que el nombre correcto es `wbc`.

De manera similar, dentro del dataset se encuentra de la variable `rc`, pero en la descripción aparece el nombre a `rbcc`. Buscando las siglas en Inglés se encontró que el nombre correcto es `rbc`.

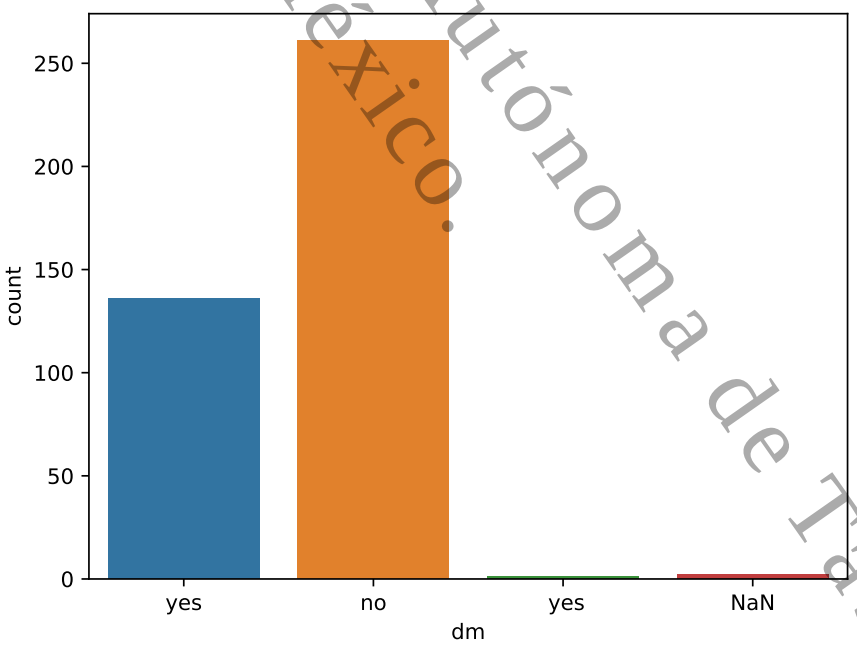

Figura 2.14. Distribución de frecuencias de la variable diabetes mellitus (`dm`).

2.7. Conclusión

En este capítulo se realizó un análisis univariado del dataset KCD. Se encontraron detalles significativos en varias variable del dataset en particular de la variable `dm` y se puede concluir que al analizar las variables tanto numéricas como nominales del dataset se tienen muchos valores nulos, ya que todas las variables tienen al menos 1 valor faltante. Por lo tanto, para futuros trabajos que impliquen la creación de modelos predictivos que ayuden a diagnosticar que persona puede tener o no la enfermedad, se debe realizar una imputación de datos en todo el dataset. Para las variables numéricas se propone reemplazar los valores faltantes con la media o mediana, y para las nominales se propone la moda. Los valores faltantes pueden disminuir la precisión un modelo o llevar a resultados sesgados, Por último, los valores atípicos puede dar lugar a estimaciones erróneas en el modelo, se propone para eliminarlos hacer una normalización de los datos.

Referencias del capítulo

Arias Rodríguez, M., Aljama García, P., Egido de los Ríos, J., Lamas Peláez, S., Praga Terente, M., & Serón, D. (2013). *Nefrología clínica*. Editorial Médica Panamericana.

Deitel, P., & Deitel, H. (2019). *Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and the Cloud* (1.ª ed.). Pearson.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3.ª ed.). O'Reilly Media.

Instituto Mexicano del Seguro Social. (2013). *Acuerdo de cooperación técnica IMSS- OPS/OMS* [Disponible en www.imss.gob.mx/NR/rdonlyres/…/convenio_cooperacion.pdf (fecha de consulta 03/06/13)].

Levey, A. S., Atkins, R., Coresh, J., Cohen, E. P., Collins, A. J., Eckardt, K. U., Nahas, M. E., Jaber, B. L., Jadoul, M., Levin, A., Powe, N. R., Rossert, J., Wheeler, D. C., Lameire, N., & Eknoyan, G. (2007). Chronic kidney disease as a global public health problem: Approaches and initiatives – a position statement from Kidney Disease Improving Global Outcomes. *Kidney International*, *72*(3), 247-259. https://doi.org/10.1038/sj.ki.5002343

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2014). *Introducción a la Probabilidad y Estadística* (14.ª ed.). CENGAGE Learning.

Myers, R. H., Walpole, R. E., & Myers, S. L. (2012). *Probabilidad y Estadística para ingeniería y ciencias* (9.ª ed.). PEARSON EDUCACIÓN.

Organización Panamericana de la Salud. (2026). *ENLACE: Portal de Datos sobre Enfermedades No Transmisibles, Salud Mental, y Causas Externas* [Accedido en 2026]. https://www.paho.org/es/enlace/carga-enfermedes-renales

Capítulo 3

Algoritmo EBM para la detección temprana de enfermedad cardiaovascular

Inteligencia artificial interpretable para la detección temprana de enfermedad

Simón Javier Hernández-Gaspar; Oscar Chávez-Bosquez

Resumen La predicción temprana de enfermedades cardiovasculares es un reto crítico en la medicina moderna, donde la precisión del modelo debe ir acompañada de una explicación clara de sus decisiones. Este artículo presenta el desarrollo de un modelo de aprendizaje automático supervisado utilizando el algoritmo Explainable Boosting Machine (EBM) aplicado a un conjunto de datos clínicos de 303 pacientes. A diferencia de los modelos de "caja negra", EBM permite una interpretación detallada tanto global como local de las predicciones. Los resultados demuestran que es posible obtener un alto rendimiento predictivo manteniendo una transparencia total sobre cómo variables como la edad, el tipo de dolor torácico y los niveles de colesterol contribuyen al riesgo individual de cada paciente. El uso de la librería interpret en Python facilita esta transición hacia una inteligencia artificial más ética y comprensible para el sector salud.

Palabras clave: Aprendizaje automático, EBM, HEART dataset.

Institución de adscripción: División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México.

3.1. Introducción

Las enfermedades cardiovasculares (ECV) constituyen un grupo de trastornos que comprometen el corazón y el sistema vascular, tanto arterias como venas. Su origen suele ser multifactorial, vinculado a determinantes socioeconómicos, conductuales y ambientales. Entre los principales factores de riesgo se incluyen la hipertensión arterial, la alimentación poco saludable, el hipercolesterolemia, la diabetes, la exposición a contaminación ambiental, la obesidad, el tabaquismo, la enfermedad renal crónica, la inactividad física, el consumo excesivo de alcohol y el estrés. Asimismo, variables no modificables como los antecedentes familiares, la edad, el sexo biológico y el origen étnico influyen de manera significativa en la probabilidad de desarrollar ECV (Federación Mundial del Corazón, 2026).

Las ECV constituyen la primera causa de mortalidad tanto en México como a nivel mundial. Patologías como el infarto agudo al miocardio y los accidentes cerebrovasculares ocasionan más de 17 millones de muertes cada año, cifra que podría incrementarse a 23.6 millones hacia 2030, de acuerdo con estimaciones de la Organización Mundial de la Salud (Organización Mundial de la Salud, 2021).

En el contexto nacional se reportan más de 150 mil defunciones anuales relacionadas con problemas cardíacos, principalmente por infarto agudo al miocardio. Esta situación ha sido descrita como una “pandemia permanente”, distinta de las epidemias infecciosas como la COVID-19, ya que la arterioesclerosis y las ECV persisten en el tiempo. Además, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) (Organización para la Cooperación y el Desarrollo Económicos, s.f.) señala que México se encuentra entre los países con mayor avance en la incidencia de estas enfermedades, mientras que otras naciones han logrado reducir su impacto mediante programas de detección, prevención y tratamiento oportuno (Dirección General de Comunicación Social, UNAM, 2020).

Dada la relevancia clínica de la enfermedad cardiovascular (ECV), resulta fundamental comprender su comportamiento para diseñar estrategias de prevención y, en caso necesario, ofrecer tratamientos adecuados a las personas con alto riesgo de desarrollarla. En este contexto, el Aprendizaje automático, rama de la Inteligencia Artificial, permite que los sistemas mejoren de manera autónoma mediante algoritmos que procesan grandes volúmenes de datos sin necesidad

de programación explícita. Estas aplicaciones automatizan la construcción de modelos estadísticos, identifican patrones y generan decisiones con mínima intervención humana (Bagnato, 2020).

Los algoritmos de aprendizaje automático pueden clasificarse en distintas familias según el área de aplicación o los resultados esperados. Existen enfoques basados en reglas, redes neuronales o métodos de optimización, entre otros. Una clasificación adicional se relaciona con su grado de interpretabilidad: los modelos de "caja gris" o *glassbox* (como árboles de decisión o regresión lineal) son inherentemente explicativos, mientras que los modelos de "caja negra" o *blackbox* (como redes neuronales profundas o bosques aleatorios) generan resultados cuya lógica interna es difícil de comprender (Molnar, 2024).

Uno de los principales retos de estos modelos es precisamente su carácter de caja negra, lo que dificulta la interpretación de sus predicciones. Por ello, se ha enfatizado la necesidad de desarrollar metodologías que otorguen transparencia y confianza a los usuarios. En este sentido, herramientas de código abierto como InterpretML ofrecen modelos interpretables, entre ellos el Explainable Boosting Machine (EBM), que permiten explicar de manera clara cómo se generan las predicciones de manera local y global.

3.2. Metodología

Para el desarrollo de este estudio se utilizó un enfoque cuantitativo y experimental basado en las etapas que se describen a continuación.

Se utilizó el conjunto de datos de enfermedades cardíacas de la Cleveland Clinic Foundation, disponible en el repositorio de Aprendizaje automático de la Universidad de California en Irvine (UCI). Aunque la base de datos original contiene hasta 76 atributos, en este estudio se empleó el subconjunto procesado ampliamente utilizado en la literatura, compuesto por 303 instancias y 14 variables clínicas (Detrano et al., 1989).

Dicho subconjunto corresponde a la selección validada por (Detrano et al., 1989), la cual ha sido adoptada como referencia estándar en estudios de predicción de enfermedad coronaria mediante técnicas de Aprendizaje automático. Las variables incluidas abarcan características demográficas (como edad y sexo) y mediciones clínicas relevantes (como presión arterial, niveles de colesterol y resultados de electrocardiograma). La variable objetivo es de tipo binario e indica la

presencia o ausencia de la enfermedad. Para efectos de implementación, el conjunto de datos fue obtenido a través de un repositorio público en GitHub, el cual contiene una versión preprocesada del dataset original de UCI, facilitando su uso en entornos computacionales (Dery, 2026).

A continuación, en la Tabla 3.1 se describen detalladamente los atributos del dataset y sus rangos de valores observados.

Tabla 3.1. Descripción de las variables del conjunto de datos de enfermedades cardíacas.

No.	Variable	Significado	Valores/Unidades
1	`age`	Edad	Años (29 - 77)
2	`sex`	Sexo	(1 = masculino, 0 = femenino)
3	`cp`	Tipo de dolor torácico	(0, 1, 2, 3)
4	`trestbps`	Presión arterial en reposo	mm/Hg
5	`chol`	Colesterol sérico	mg/dl
6	`fbs`	Azúcar en sangre en ayunas	(¿120 mg/dl: 1, $\leq$ 120 mg/dl: 0)
7	`restecg`	Resultados electrocardiográficos	(0, 1, 2)
8	`thalach`	Frecuencia cardíaca máxima	Latidos por minuto
9	`exang`	Angina inducida por ejercicio	(1 = sí, 0 = no)
10	`oldpeak`	Depresión del ST (ejercicio/reposo)	Valor numérico (0.0 - 6.2)
11	`slope`	Pendiente del segmento ST	(0, 1, 2)
12	`ca`	Vasos coloreados por fluoroscopia	(0, 1, 2, 3, 4)
13	`thal`	Tipo de talasemia	(0, 1, 2, 3)
14	`target`	Diagnóstico (variable objetivo)	(0 = sin enfermedad, 1 = enfermo)

Para desarrollar el análisis y modelado de los datos se utilizó la biblioteca Scikit-learn de Python, ya que permite tanto el aprendizaje supervisado como el no supervisado y es de código abierto (Pedregosa et al., 2011). En particular, se utilizó la biblioteca Pandas para manipular y analizar los datos, así como las bibliotecas Matplotlib y Seaborn. Para visualizar cada una de las variables Se clasificaron las variables en dos categorías para optimizar el entrenamiento del modelo:

- Variables continuas: edad, presión arterial, colesterol, frecuencia cardíaca máxima y depresión del ST.
- Variables categóricas: sexo, tipo de dolor torácico, azúcar en la sangre, resultados electrocardiográficos, entre otras.

3.2.1. Algoritmos interpretables

La incorporación de la Inteligencia Artificial en el ámbito de la salud requiere mantener un equilibrio entre la precisión diagnóstica y la transparencia de los modelos. En este trabajo, la *interpretabilidad* se entiende como la capacidad de comprender la relación lógica entre las variables que intervienen en el modelo (Molnar, 2024), mientras que la *explicabilidad* se refiere a la posibilidad de ofrecer justificaciones claras sobre las decisiones generadas por el sistema (Arrieta et al., 2020).

El Explainable Boosting Machine (EBM) es un método de Aprendizaje automático basado en modelos aditivos generalizados (GAMs). A diferencia de los modelos lineales tradicionales, el EBM captura relaciones no lineales mediante funciones de forma (f_j) para cada característica, manteniendo una transparencia total Ganie et al., 2025. Su formulación matemática se define como:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j) \tag{3.1}$$

Donde g es la función de enlace, β_0 el intercepto, f_j las funciones de efectos principales y f_{ij} las interacciones por pares.

Contribuciones en EBM

En la clasificación binaria de riesgo, EBM utiliza la función de enlace *logit*, operando en el espacio de *log-odds*. La probabilidad final (P) de pertenecer a la clase de riesgo (Clase 1) se obtiene mediante la suma aditiva de las contribuciones, transformada por la función logística inversa:

$$P(y = 1|x) = \frac{1}{1 + \exp(-\text{score})} \tag{3.2}$$

Esta estructura aditiva justifica matemáticamente la interpretación clínica de las explicaciones locales:

- Valores positivos (barras naranjas): Cuando la función da como resultado $f_j(x_j) > 0$, la característica aumenta el valor del *log-odds*. Matemáticamente, esto empuja la probabilidad

P hacia 1, lo que se traduce clínicamente en un aumento directo del riesgo estimado de desarrollar la complicación.

- Valores negativos (barras azules): Cuando $f_j(x_j) < 0$, la característica reduce el valor del *log-odds*. Esto empuja la probabilidad P hacia 0, actuando clínicamente como factor protector o indicador de salud, desplazando la predicción final hacia la ausencia de complicaciones (Clase 0).

De este modo, cada variable impacta de forma independiente y exacta en el logaritmo de la probabilidad, permitiendo identificar con precisión qué factores suman o restan riesgo para cada paciente.

División de los datos y validación del modelo

Para el desarrollo del modelo, el conjunto de datos fue dividido en variables predictoras (X) y variable objetivo (y). Posteriormente, se realizó una partición del conjunto de datos en subconjuntos de entrenamiento y prueba utilizando la función `train_test_split` de la biblioteca *scikit-learn* (Pedregosa et al., 2011).

Se asignó el 80 % de los datos al conjunto de entrenamiento y el 20 % al conjunto de prueba (`test_size = 0.2`). Esta estrategia permite entrenar el modelo con una proporción representativa de los datos disponibles y evaluar su capacidad de generalización sobre datos no vistos. Asimismo, se estableció un valor fijo de semilla aleatoria (`random_state = 42`) con el objetivo de garantizar la reproducibilidad de los resultados.

El modelo basado en EBM requiere la especificación del tipo de variables de entrada. En este estudio, las variables fueron clasificadas como continuas y categóricas (nominales), permitiendo al modelo ajustar funciones específicas según la naturaleza de cada variable.

La evaluación del modelo se llevó a cabo utilizando el conjunto de prueba, el cual no fue empleado durante el entrenamiento. Este enfoque, conocido como validación *hold-out*, permite estimar el desempeño del modelo en datos independientes y reduce el riesgo de sobreajuste. Con el objetivo de evaluar la estabilidad del modelo, se complementó la estrategia de validación *hold-out* con un esquema de validación cruzada estratificada de 5 particiones (*Stratified k-fold cross-validation*). Este procedimiento permitió estimar la variabilidad de las métricas de desempeño

a través de diferentes particiones de los datos. A partir de los resultados obtenidos en cada partición, se calcularon intervalos de confianza del 95 %, proporcionando una estimación robusta del rendimiento del modelo.

Adicionalmente, se realizó una prueba t de una muestra para evaluar si el desempeño del modelo es significativamente superior al azar. Los resultados indicaron significancia estadística ($p < 0.05$), lo que respalda la validez de las predicciones obtenidas.

Finalmente, para el análisis interpretativo del modelo, se seleccionaron instancias del conjunto de prueba y se generaron explicaciones locales mediante la función `explain_local`, lo que permitió identificar la contribución de cada variable en las predicciones individuales (Molnar, 2024).

3.3. Estado del arte

La literatura reciente destaca una transición desde modelos de Aprendizaje automático convencionales hacia arquitecturas de Inteligencia Artificial Explicable (XAI). Diversos estudios han validado el uso del conjunto de datos de la Cleveland Clinic (UCI Heart Disease) como un estándar para evaluar la interpretabilidad en diagnósticos clínicos. Investigaciones previas han demostrado que, aunque algoritmos de "caja negra" como las redes neuronales ofrecen una alta precisión, la capacidad de EBM para descomponer el riesgo en contribuciones individuales por variable resulta superior para la toma de decisiones médicas, permitiendo identificar factores críticos como el tipo de dolor torácico y la frecuencia cardíaca máxima sin perder transparencia diagnóstica. La Tabla 6.2 resume el estado del arte.

La predicción de enfermedades cardiovasculares ha evolucionado desde los modelos estadísticos tradicionales hasta los enfoques actuales de Inteligencia Artificial Explicable (XAI). El trabajo pionero de (Detrano et al., 1989), puede considerarse el "artículo semilla", ya que fue donde se validó el conjunto de 14 variables clínicas que hoy se consideran el estándar para este propósito. Recientemente, investigaciones como la de (Ganie et al., 2025) han demostrado que el uso de ensambles de Aprendizaje automático, combinados con técnicas de explicabilidad, no solo mejora la precisión diagnóstica, sino que proporciona la transparencia necesaria para su adopción en entornos clínicos reales. Asi también (Agrawal et al., 2025) propone un marco de trabajo que combina el rendimiento de modelos complejos con la transparencia de XAI, facilitando

Tabla 3.2. Trabajos relacionados con este estudio.

Referencia	Método	Aporte principal
Detrano et al., 1989	Regresión Logística	Validación clínica de las 14 variables originales de Cleveland Clinic.
Ganie et al., 2025	Ensemble + SHAP	Mejora diagnóstica con enfoque en transparencia clínica.
Paul, 2024	LIME + SHAP	Uso de técnicas de interpretabilidad local y global para explicar modelos de IA en sistemas médicos.
Hernández-Gaspar et al., 2025	EBM	Aplicación de modelos interpretables para el análisis de pacientes con insuficiencia renal crónica.
Agrawal et al., 2025	XGBoost + LIME + SHAP	Integración de múltiples algoritmos con capas de interpretabilidad para predicción de diversas patologías hospitalarias.
Yagin et al., 2025	EBM	Uso de modelos intrínsecamente interpretables para identificar biomarcadores de sepsis en tiempo real en urgencias.

la toma de decisiones clínicas en patologías, cubriendo la necesidad de sistemas que los médicos puedan auditar visualmente. Y como última referencia (Yagin et al., 2025) hace un estudio particularmente relevante porque utiliza EBM en un entorno de alta presión como es el área de urgencias. Demuestra que este algoritmo no solo iguala la precisión de modelos de caja negra, sino que permite descubrir relaciones no lineales entre los signos vitales y el riesgo de sepsis.

3.4. Resultados

En esta sección se presentan los hallazgos obtenidos tras el entrenamiento y validación del modelo EBM. El desempeño se evaluó no solo desde una perspectiva cuantitativa, sino también mediante la capacidad del modelo para discernir entre clases en un entorno clínico.

En la figura 3.1 se muestra parte del código usado para generar las medidas de rendimiento, la matriz de confusión, así como la matriz de correlación.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.metrics import (confusion_matrix, ConfusionMatrixDisplay,
                             accuracy_score, precision_score, recall_score,
                             f1_score, classification_report)
from interpret.glassbox import ExplainableBoostingClassifier
from sklearn.model_selection import train_test_split

# 1. Cargar el dataset
df = pd.read_csv('heart.csv')

# 2. Entrenamiento de EBM
X = df.drop('target', axis=1)
y = df['target']
feature_types = ['continuous', 'nominal', 'nominal', 'continuous', 'continuous',
                 'nominal', 'nominal', 'continuous', 'nominal', 'continuous',
                 'nominal', 'nominal', 'nominal']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
ebm = ExplainableBoostingClassifier(feature_types=feature_types, random_state=42)
ebm.fit(X_train, y_train)

# 3. Predicciones y Métricas
y_pred = ebm.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
report = classification_report(y_test, y_pred, target_names=['Sano', 'Enfermo'])
```

Figura 3.1. Código para generar medidas de rendimiento, matriz de correlación y matriz de confusión.

3.4.1. Evaluación del desempeño del modelo

Para observar la distribución de las predicciones se generó la matríz de confusión sobre el conjunto de prueba, la cual permite identificar el comportamiento del algoritmo frente a los casos positivos y negativos (Figura 3.2). Como se observa en la Figura, de un total de 61 pacientes evaluados:

- 25 pacientes fueron identificados correctamente como sanos (verdaderos negativos).
- 27 pacientes fueron identificados correctamente con presencia de enfermedad (verdaderos positivos).
- Se registraron 5 falsos positivos, lo que representa casos donde el modelo sugirió riesgo en ausencia de la patología.
- Se registraron 4 falsos negativos, que corresponden a pacientes con enfermedad que el modelo no logró detectar.

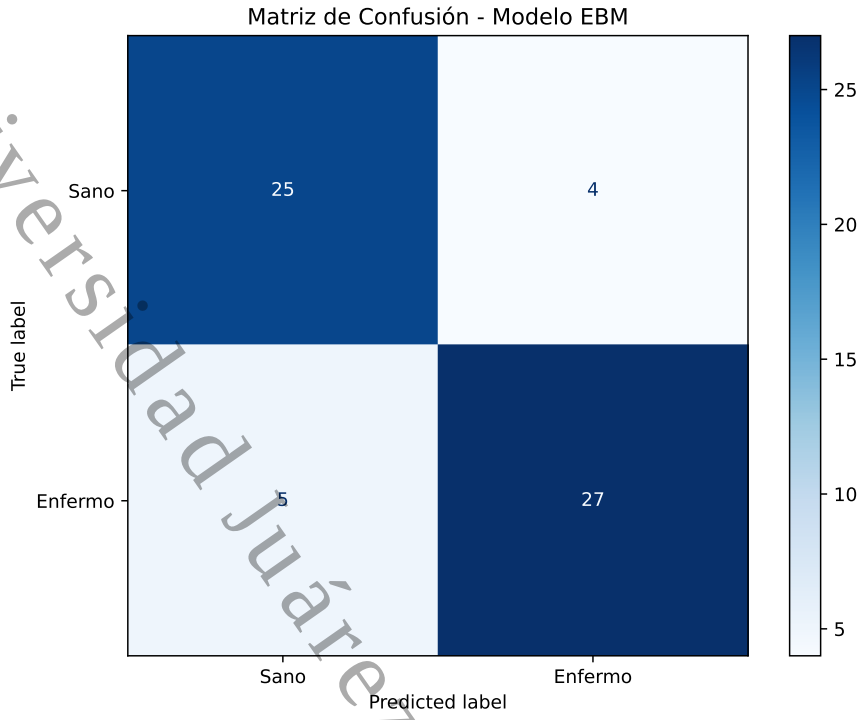

Figura 3.2. Matriz de confusión del modelo EBM para el diagnóstico de enfermedad cardiaca.

A partir de los resultados de esta matriz, se calcularon las métricas de rendimiento que se resumen en la Tabla 3.3.

Tabla 3.3. Métricas de desempeño del modelo EBM.

Métrica	**Valor**
Exactitud (*Accuracy*)	0.8525
Precisión (*Precision*)	0.8710
Sensibilidad (*Recall*)	0.8438
Puntuación F1 (*F1-Score*)	0.8571

3.4.2. Análisis del desempeño

El modelo logra una exactitud (*Accuracy*) del 85.25 %, lo que indica una capacidad predictiva aceptable sobre el conjunto de datos de prueba, clasificando correctamente a la gran mayoría de los pacientes, tanto sanos como enfermos.

Con respecto a la sensibilidad (*Recall*), considerada la métrica más importante en este estudio clínico, se alcanzó un valor de 0.8438. Esto significa que el modelo es capaz de identificar

correctamente al 84.38 % de los pacientes que realmente padecen una afección cardíaca. En el contexto médico, una sensibilidad elevada es fundamental, ya que el objetivo primordial es minimizar los falsos negativos, asegurando que la mayoría de los individuos en riesgo reciban la atención y el seguimiento necesarios.

Finalmente, el balance entre la precisión (0.8710) y la sensibilidad (0.8438) se ve reflejado en un *F1-Score* de 0.8571, lo que demuestra que el modelo mantiene un equilibrio robusto para detectar la enfermedad de manera confiable.

3.4.3. Matriz de correlación

Para identificar las variables clínicas con mayor influencia sobre el diagnóstico, se realizó un análisis de correlación de Pearson entre cada atributo y la variable objetivo (`target`). Los resultados, pueden verse en la Figura 3.3, la cual es detallada en la Tabla 3.4, permitiendo jerarquizar los factores de riesgo según su fuerza de asociación.

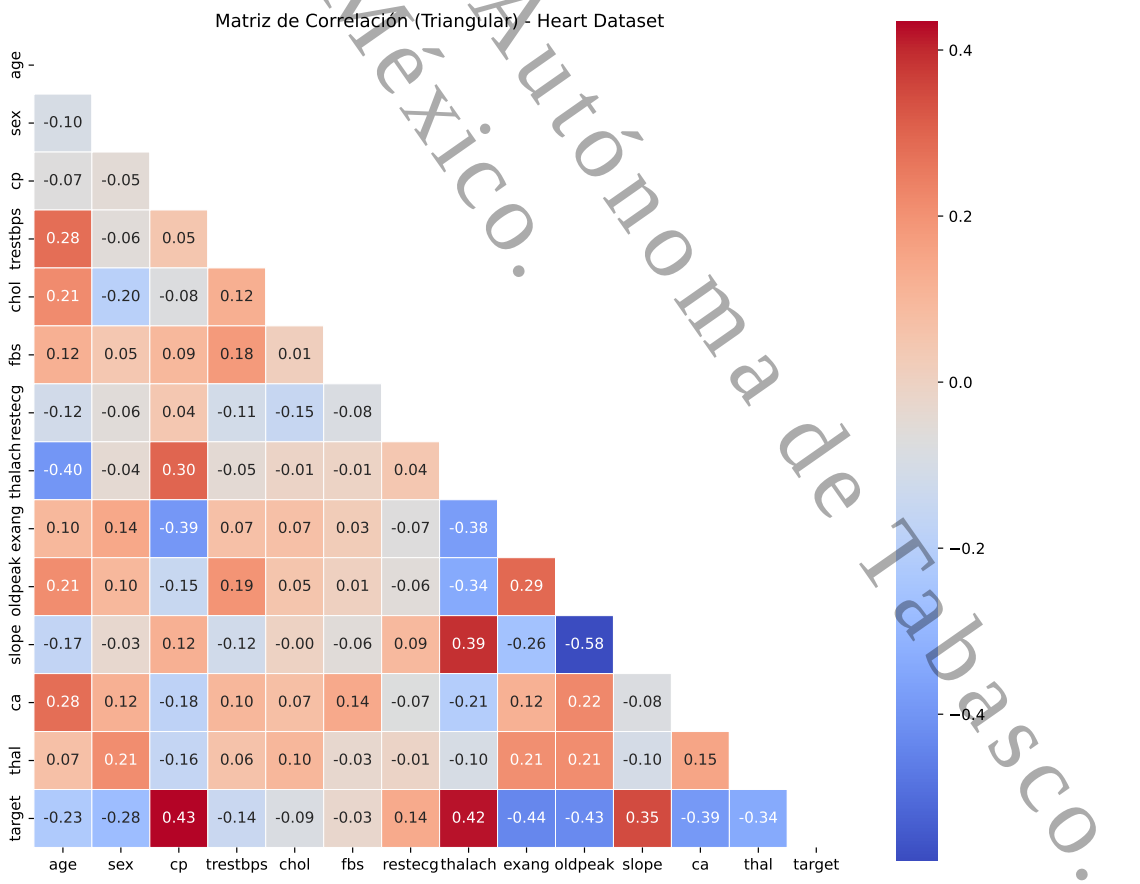

Figura 3.3. Matriz de correlación de las variables con la `target`.

Tabla 3.4. Coeficientes de correlación de las variables con la variable objetivo.

Variable	Coeficiente de Correlación	Tipo de Relación
`cp` (Dolor torácico)	0.4338	Directa (Fuerte)
`thalach` (Frecuencia cardíaca máx.)	0.4217	Directa (Fuerte)
`slope` (Pendiente ST)	0.3459	Directa
`restecg` (Electrocardiograma)	0.1372	Directa (Débil)
`fbs` (Azúcar en sangre)	-0.0280	Despreciable
`chol` (Colesterol)	-0.0852	Inversa (Débil)
`trestbps` (Presión arterial)	-0.1449	Inversa (Débil)
`age` (Edad)	-0.2254	Inversa
`sex` (Sexo)	-0.2809	Inversa
`thal` (Talasemia)	-0.3440	Inversa
`ca` (Vasos coloreados)	-0.3917	Inversa (Fuerte)
`oldpeak` (Depresión ST)	-0.4307	Inversa (Fuerte)
`exang` (Angina inducida)	-0.4368	Inversa (Fuerte)

De la matriz de correlación se obtienen los siguientes puntos:

- Las variables `cp` (tipo de dolor de pecho) y *thalach* (frecuencia cardíaca máxima alcanzada) muestran las correlaciones positivas más altas (0.43 y 0.42 respectivamente). Esto indica que los niveles altos de estos indicadores están asociados significativamente con la presencia de enfermedad cardíaca.
- Se observa una fuerte correlación inversa en variables como `exang` (-0.43) y `oldpeak` (-0.43). Esto sugiere que la presencia de angina inducida por ejercicio y mayores niveles de depresión del segmento ST actúan como indicadores inversos de salud en la clasificación del modelo.
- Es importante mencionar que variables críticas como el `fbs` (azúcar en ayunas) y el `chol` (colesterol) presentan coeficientes cercanos a cero (-0.02 y -0.08), lo que sugiere que su relación con la variable objetivo no es lineal o requiere de la interacción con otros factores para ser significativa.

3.4.4. Validación estadística del modelo

Con el objetivo de evaluar la estabilidad y robustez del modelo, se implementó un esquema de validación cruzada estratificada de 5 particiones utilizando la biblioteca *scikit-learn*. Los valores de

exactitud obtenidos en cada partición fueron los siguientes: $[0.9344, 0.8689, 0.7541, 0.8500, 0.8333]$. A partir de estos resultados, se obtuvo una *Accuracy* promedio de 0.8481 ± 0.0590, lo que sugiere un desempeño consistente del modelo a través de diferentes subconjuntos de los datos.

Asimismo, se calculó un intervalo de confianza (IC) del 95 % para la exactitud: $IC_{95\,\%} = [0.7971, 0.8992]$. Este intervalo indica que el desempeño del modelo se mantiene dentro de un rango estable, lo que refuerza su capacidad de generalización. Finalmente, se realizó una prueba t de una muestra para evaluar si el desempeño del modelo es significativamente superior al azar. Los resultados obtenidos indican que el modelo presenta una mejora estadísticamente significativa ($p < 0.05$), lo que respalda la validez de sus predicciones.

3.4.5. Interpretación del Modelo EBM: Importancia global de los atributos

Una de las ventajas de *Explainable Boosting Machine* (EBM) es que permite analizar visualmente la relación entre la correlación estadística y la importancia de las variables. Mientras que la correlación (Tabla 3.4) mide relaciones lineales, el análisis global de EBM captura la contribución de cada atributo al proceso de decisión del modelo.

En la Figura 3.4 se muestra parte del código usado para generar la gráfica global de EBM; mediante este código se genera el gráfico de la importancia de cada instancia con referencia a la variable objetivo.

```
# Definimos tipos (EBM diferencia continuas de categóricas)
feature_types = ['continuous', 'nominal', 'nominal', 'continuous', 'continuous',
                 'nominal', 'nominal', 'continuous', 'nominal', 'continuous',
                 'nominal', 'nominal', 'nominal']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 2. Entrenar EBM
ebm = ExplainableBoostingClassifier(feature_types=feature_types)
ebm.fit(X_train, y_train)

# 3. Seleccionar 2 pacientes del test para explicar
X_sample = X_test.head(2)
ebm_local = ebm.explain_local(X_sample, y_test.head(2))


# Esto te mostrará cómo el modelo "ve" cada variable en general
ebm_global = ebm.explain_global()
show(ebm_global)
```

Figura 3.4. Código para generar gráfica global de EBM.

En la Figura 3.5 se presenta la importancia global de las variables. A diferencia de la matriz

de correlación, este gráfico pondera cuánto contribuye cada característica al error del modelo en promedio.

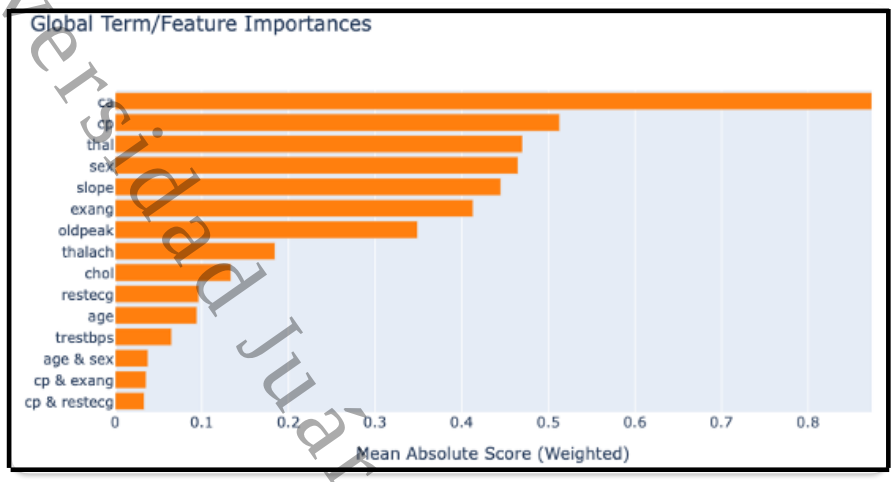

Figura 3.5. Importancia global de las variables según el modelo EBM.

Como se puede observar, las variables `cp` (dolor torácico), `thal` (talasemia) y `ca` (número de vasos coloreados) se posicionan como los predictores con mayor peso en el modelo resultante.

3.4.6. Interpretación del Modelo EBM: Análisis de casos individuales (explicación local)

Se muestra a continuación en la Figura 3.5 el código que genera las gráficas de las predicciones locales obtenidas por el algoritmo EMB. Luego detallamos 2 casos, uno donde el modelo hace una predicción correcta y otra donde se equivoca, tomando el caso más severo para el análisis clínico.

Caso 1: Predicción correcta (verdadero positivo)

En este escenario, el modelo EBM identifica correctamente a un paciente con presencia de enfermedad cardíaca, asignando una probabilidad del 90.6 %. El *score logit* acumulado es de 2.261. Los factores de riesgo predominantes son:

```python
# Definimos tipos (EBM diferencia continuas de categóricas)
feature_types = ['continuous', 'nominal', 'nominal', 'continuous', 'continuous',
                 'nominal', 'nominal', 'continuous', 'nominal', 'continuous',
                 'nominal', 'nominal', 'nominal']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 2. Entrenar EBM
ebm = ExplainableBoostingClassifier(feature_types=feature_types)
ebm.fit(X_train, y_train)

# 3. Seleccionar 5 pacientes del test para explicar
X_sample = X_test.head(5)
ebm_local = ebm.explain_local(X_sample, y_test.head(5))

# Para ver las gráficas de los 5 pacientes:
from interpret import show
show(ebm local)
```

Figura 3.6. Código para generar predicciones locales.

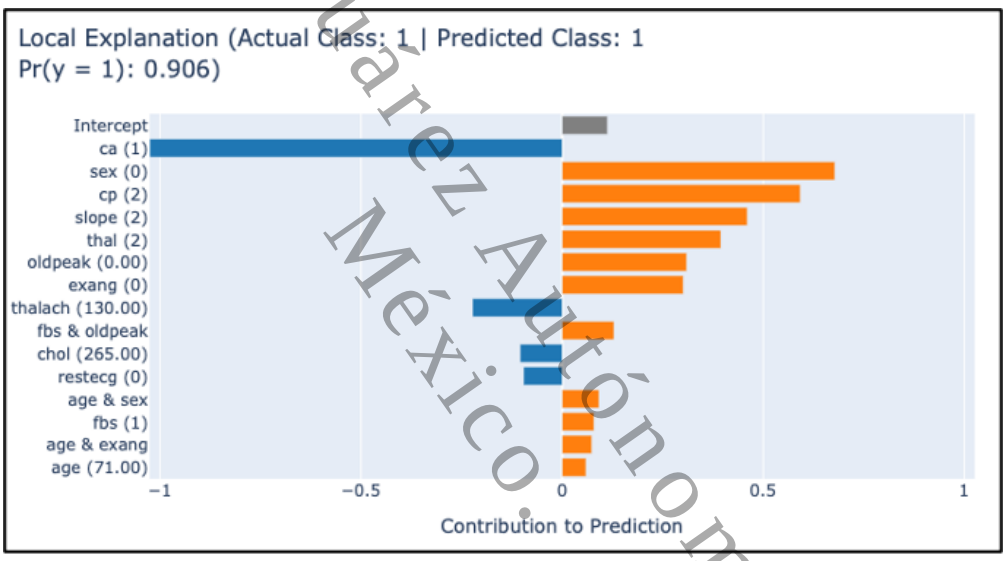

Figura 3.7. Predicción correcta de un paciente de EBM.

- `cp = 2.000`: El tipo de dolor torácico es el principal contribuyente, incrementando significativamente el *score* hacia la clase positiva.
- `ca = 1.000`: La presencia de vasos coloreados contribuye positivamente al riesgo en este paciente.
- `slope` y `thal`: También aportan de manera relevante al incremento del *score*.

En contrapartida, los factores protectores son:

- `thalach = 130.000`: Contribuye negativamente al modelo, reduciendo ligeramente la probabilidad de enfermedad.

Como factor adicional se tiene:

- `age = 71.000`: Presenta una contribución positiva menor en comparación con otras variables.

El modelo parte de un intercepto (β_0) de 0.678, al cual se suman las contribuciones locales (f_i) resultando en un *score* total de 2.261. Al aplicar la función logística, se obtiene una probabilidad de $P = 0.906$, lo que confirma una predicción correcta con alta confianza hacia la Clase 1.

Caso 2: Predicción Incorrecta (falso negativo)

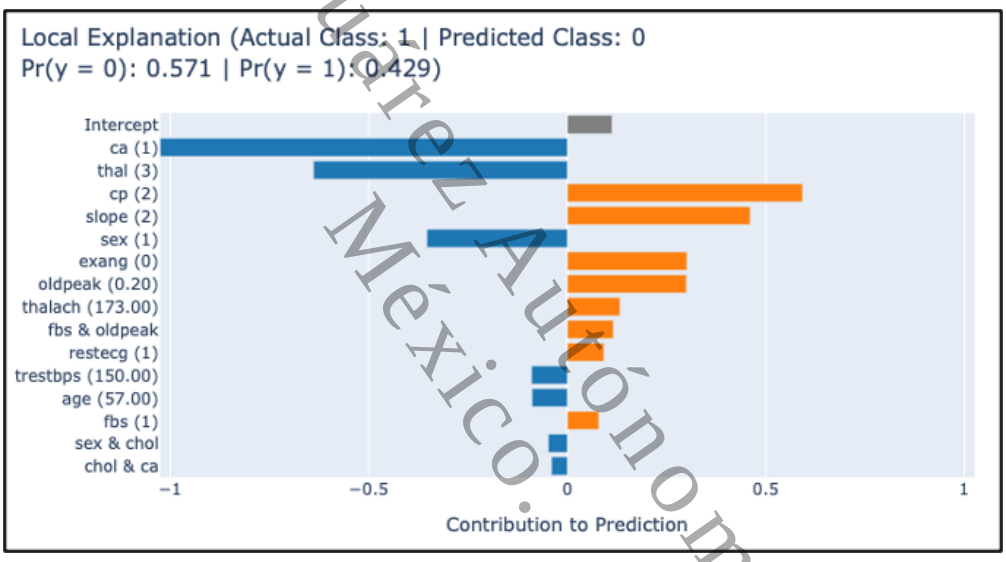

Figura 3.8. Predicción incorrecta de un paciente de EBM.

En este escenario, el modelo EBM clasifica erróneamente a un paciente con enfermedad cardíaca (Clase real = 1) como sano (Clase predicha = 0), con una probabilidad de 0.571 para la clase negativa. Este tipo de error es clínicamente importante, ya que implica la omisión de un diagnóstico positivo.

En este caso, los factores se distribuyen de la siguiente manera:

- Factores que favorecen la clase positiva: variables como `cp = 2`, `slope = 2`, `exang = 0` y `oldpeak = 0.20` contribuyen positivamente al *score*, sugiriendo la presencia de enfermedad.

- Factores dominantes hacia la clase negativa: Variables como `ca = 1`, `thal = 3` y `sex = 1` presentan contribuciones negativas de mayor magnitud, dominando el comportamiento del modelo y reduciendo la probabilidad final de la clase positiva.

Este caso evidencia un conflicto entre variables clínicas relevantes y patrones aprendidos por el modelo, donde las contribuciones negativas superan a las positivas, resultando en una subestimación del riesgo real del paciente. Este tipo de error resalta la importancia de los mecanismos interpretables, ya que permiten identificar situaciones donde el modelo puede fallar debido a la interacción o predominancia de ciertas variables.

3.5. Discución y conclusiones

El modelo basado en EBM demostró ser una herramienta eficaz para la clasificación de enfermedad cardíaca, alcanzando un desempeño equilibrado entre precisión y sensibilidad. El principal aporte del modelo radica en su capacidad interpretativa, ya que permite analizar tanto la importancia global de las variables como su contribución individual en cada predicción, facilitando una comprensión más profunda del comportamiento del modelo.

El modelo presenta un equilibrio adecuado entre precisión (0.8710) y sensibilidad (0.8438), lo cual es especialmente relevante en aplicaciones médicas, donde la detección oportuna de la enfermedad es prioritaria. No obstante, el análisis de la matriz de confusión evidencia la presencia de falsos negativos, los cuales representan un riesgo clínico significativo al implicar la omisión de pacientes con patología.

Adicionalmente, los resultados obtenidos mediante validación cruzada mostraron un desempeño consistente del modelo, con una exactitud promedio de 0.8481 y un intervalo de confianza del 95 % que indica estabilidad en su capacidad de generalización. Asimismo, la prueba de significancia estadística confirmó que el desempeño del modelo es superior al azar ($p < 0.05$), lo que respalda la validez de los resultados obtenidos.

El análisis de casos individuales, particularmente los falsos negativos, puso de manifiesto que los errores del modelo surgen de la interacción y competencia entre variables, lo que puede conducir a una subestimación del riesgo en pacientes con señales clínicas relevantes. Este hallazgo

refuerza la importancia de incorporar modelos interpretables en aplicaciones médicas, donde no solo es necesario predecir con precisión, sino también comprender las razones detrás de cada decisión.

Finalmente, estos resultados sugieren que el uso de modelos explicables como EBM no solo mejora la transparencia, sino que también ofrece herramientas valiosas para la validación clínica, la identificación de sesgos y la toma de decisiones informadas en sistemas de apoyo al diagnóstico.

Referencias del capítulo

Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology*, *20*(1), 105. https://doi.org/10.1186/s13000-025-01686-3

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bagnato, J. I. (2020). Aprende Machine Learning en Español: Teoría + Práctica Python. https://www.aprendemachinelearning.com

Dery, L. (2026). Heart disease dataset (heart.csv) [Dataset disponible en el repositorio EDA-course]. https://github.com/nlihin/EDA-course/blob/main/datasets/heart.csv

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9

Dirección General de Comunicación Social, UNAM. (2020, septiembre). Enfermedades del corazón, pandemia permanente.

Federación Mundial del Corazón. (2026). ¿Qué son las enfermedades cardiovasculares? [Accedido: 08-abr-2026].

Ganie, S. M., Pramanik, P. K. D., & Zhao, Z. (2025). Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Scientific Reports*, *15*(13912). https://doi.org/10.1038/s41598-025-97547-6

Hernández-Gaspar, S. J., Hernández-Ocaña, B., Hernández-Torruco, J., & Chávez-Bosquez, O. (2025). Inteligencia artificial explicable para el análisis de pacientes con insuficiencia renal crónica. *Nova Scientia*, *17*, 1-22. https://doi.org/10.21640/ns.v17iSpecialIa.3630

Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book

Organización Mundial de la Salud. (2021, junio). Enfermedades cardiovasculares.

Organización para la Cooperación y el Desarrollo Económicos. (s.f.). Acerca de la OCDE [Accedido: 08-abr-2026].

Paul, J. (2024). LIME and SHAP interpretability for medical AI systems. https://doi.org/10.13140/RG.2.2.14652.45443

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Yagin, F. H., Aygun, U., Colak, C., Alkhalifa, A. K., Alzakari, S. A., & Aghaei, M. (2025). Accuracy is not enough: Explainable boosting machine model and identification of candidate biomarkers for real-time sepsis risk assessment in the emergency department. *BMC Emergency Medicine*, *25*(1), 248. https://doi.org/10.1186/s1471-227X-25-248

Capítulo 4

Algoritmo EBM para el análisis de pacientes con insuficiencia renal crónica

Inteligencia Artificial explicable para el análisis de pacientes con insuficiencia renal crónica

Simón Javier Hernández-Gaspar; Betania Hernández-Ocaña; José Hernández-Torruco; Oscar Chávez-Bosquez

Resumen El objetivo de este artículo es comparar algoritmos de aprendizaje automático contra un algoritmo de Inteligencia Artificial explicable en la clasificación de pacientes con enfermedad renal crónica del riñón (ERC), evaluando cuál ofrece mejores interpretaciones sin comprometer la precisión del diagnóstico. El estudio utiliza un conjunto de datos público de 400 instancias y 24 atributos para desarrollar modelos de aprendizaje automático que puedan predecir acertadamente la ERC. Se utilizaron cuatro algoritmos de aprendizaje automático: Árboles de decisión, Naïve Bayes, Máquinas de vector soporte (SVM) y Gradient boosting. Se comparó el rendimiento de estos algoritmos considerados clásicos en la literatura contra un algoritmo de aprendizaje automático explicable llamado Explainable Boosting Machine (EBM). EBM se utiliza para proporcionar una explicación de las predicciones del modelo, lo que permite entender qué variables son más importantes para la predicción. Los resultados muestran que EBM tiene un rendimiento similar al de los otros algoritmos (0.99 de *Accuracy)*, pero proporciona una explicación más clara de las predicciones. En este caso las variables más importantes para la predicción de la ERC son la gravedad específica, la hemoglobina y la creatinina sérica. El estudio concluye que el uso de modelos de aprendizaje automático explicables puede ser útil para mejorar la precisión y la transparencia en la predicción de la enfermedad renal crónica.

Palabras clave: Correlación, gráfica de coordenadas paralelas, insuficiencia renal crónica, interpretabilidad, variables predictoras.

Institución de adscripción: División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco.

4.1. Introducción

La enfermedad renal crónica del riñón (ERC), también llamada insuficiencia renal crónica, es un problema de salud pública importante, y se describe como la pérdida gradual de la función renal. La función principal de los riñones es eliminar los desechos y el exceso de líquidos de la sangre, los cuales se expulsan a través de la orina. En las etapas avanzadas de la enfermedad renal crónica, pueden acumularse en el organismo niveles perjudiciales de líquidos, electrolitos y desechos (Arias Rodríguez et al., 2013). Datos de la organización internacional World Kidney Day informan que el 10 % de la población mundial sufre de enfermedad renal crónica, la cual puede ser letal si no recibe tratamiento. Además, la mortalidad asociada a esta enfermedad incrementa anualmente (World Kidney Day, 2022).

La ERC es una enfermedad asintomática en sus primeros estadios y se estima que se convertirá en la quinta causa de muerte a nivel mundial en 2040. La tasa de mortalidad ajustada por edad debida a enfermedades renales se estimó en 15.6 defunciones por cada 100,000 habitantes en el 2019. En la mayoría de los países, la tasa de mortalidad por enfermedades renales fue mayor en hombres que en mujeres. Los países con las tasas de mortalidad (ajustadas por edad) debidas a enfermedades renales más altas en el 2019 fueron: Nicaragua, El Salvador, Bolivia, Guatemala, Surinam, Honduras y Ecuador (Organización Panamericana de la Salud, 2021).

En 2017, la prevalencia de la enfermedad renal crónica (ERC) en México fue del 12.2 %, con una tasa de 51.4 muertes por cada 100 mil habitantes. Esta enfermedad genera un impacto significativo tanto en las finanzas de las instituciones como en la economía familiar. En 2014, el costo anual promedio por paciente fue estimado en $8,966 USD en la Secretaría de Salud y en $9,091 USD en el Instituto Mexicano del Seguro Social (Instituto Nacional de Salud Pública, 2020).

De acuerdo con la Secretaría de Salud del estado de Tabasco, del año 2018 al año 2021 en Tabasco se han registrado 671 casos de insuficiencia renal. Los registros se duplicaron en el año 2020 y se triplicaron durante el año 2021, respecto al año 2019 y 2018. En ese mismo tiempo, las muertes por problemas renales han llegado a más de mil defunciones en los últimos cuatro años. Esos datos muestran que los casos de ERC se dispararon con la pandemia del COVID-19 (El Heraldo de México, 2022). Por todo lo anterior, interesa saber el comportamiento de la ERC para

intentar prevenir la afección y en dado caso proporcionar un tratamiento adecuado a las personas con alto índice de probabilidad de contraer esta enfermedad.

Por otro lado, el aprendizaje automático es una rama de la inteligencia artificial que permite que un sistema aprenda y mejore de forma autónoma mediante algoritmos sin que estos sean programados explícitamente, utilizando para ello cierta cantidad de datos. De hecho, las aplicaciones de aprendizaje automático automatizan el trabajo de compilar modelos estadísticos. El aprendizaje automático aprende de los datos, identifica patrones y toma decisiones con una intervención humana mínima (Bagnato, 2020).

Existen múltiples algoritmos que se pueden clasificar en diversas familias, dependiendo del área o de los resultados que se pretenden obtener. Hay algoritmos basados en reglas, en redes neuronales, métodos basados en optimización, por citar algunos ejemplos. Otra clasificación es de acuerdo a su interpretabilidad. Existen algoritmos que producen modelos de “caja gris” o *glassbox* (interpretables y transparentes, tales como Árboles de decisión, Regresión lineal), los cuales son llamados algoritmos inherentes, ya que son interpretables por sí mismos. Su propia estructura los hace explicativos. También existen los algoritmos que producen modelos de “caja negra” o *blackbox* (Redes neuronales, Bosques aleatorios), en los cuales no se puede saber a ciencia cierta cómo es que producen sus resultados (Molnar, 2024).

Uno de los problemas de la mayoría de los modelos de Aprendizaje automático es que se vuelven más parecidos a una caja negra, es decir, son más difíciles de entender e interpretar sus respuestas. Por eso se propone que, para obtener la confianza de las partes interesadas, es fundamental desarrollar herramientas y metodologías que ayuden al usuario a conocer y explicar cómo se realizan las predicciones. Para ello usaremos la biblioteca de código abierto InterpretML, la cual incluye modelos interpretables, siendo uno de ellos EBM (Explainable Boosting Machine).

4.2. Materiales y Métodos

4.2.1. Conjunto de datos

El dataset Chronic Kidney Disease (insuficiencia renal crónica) del repositorio de UCI (UC Machine Learning Repository), consta de 400 instancias y 25 atributos, incluyendo la clase. Se

cuenta con 14 atributos nominales y 11 atributos numéricos. En la Tabla 4.1 se describe cada uno de los atributos, con su unidad de medida y valores.

La elección de este conjunto de datos se fundamentó en varios factores que garantizan la calidad y relevancia de esta investigación. En primer lugar, este dataset cuenta con un número importante de instancias, lo que permite que los modelos de aprendizaje automático puedan generalizar los patrones y relaciones entre las variables. Un mayor número de datos contribuye a la robustez del entrenamiento y la evaluación, reduciendo la posibilidad de sesgo en los resultados.

Otro aspecto relevante es la calidad y estructura de los datos. En comparación con otros conjuntos de datos disponibles sobre insuficiencia renal, el dataset CKD de UCI presenta una estructura más completa y organizada, con menor cantidad de valores faltantes y una mejor representación de los factores clínicos asociados a la enfermedad. Esto facilita la interpretación de los resultados y mejora la precisión de los modelos utilizados. Además, este dataset ha sido ampliamente utilizado en la literatura científica, lo que permite comparar los resultados obtenidos con estudios previos y validar las conclusiones en referencia a investigaciones existentes.

De acuerdo con el análisis realizado, se observó que todos los atributos presentan valores faltantes, los cuales están representados con el carácter "?". Esto implica realizar una imputación de datos en todas las variables antes de aplicar los algoritmos de aprendizaje automático para desarrollar los modelos de predicción de la enfermedad (Bagnato, 2020).

La variable `dm` presentó particularidades en sus valores, ya que incluía espacios en blanco en algunas instancias, lo cual era contabilizado como una observación distinta. Asimismo, el rango de la mayoría de los atributos numéricos varía significativamente; se observa que la variable `wbcc` se encuentra en miles, `bgr` en cientos y `sg` en unidades, por lo que requieren un proceso de normalización.

Para el preprocesamiento de los datos, se realizaron los siguientes ajustes.

- Actualizar los valores faltantes con la media en los atributos numéricos y con el valor más frecuente (moda) en las variables nominales.
- Codificar las etiquetas para los valores de los atributos nominales.
- Escalar y transformar los datos para normalizar sus magnitudes, evitando así posibles sesgos.

Tabla 4.1. Descripción de las variables del dataset Chronic Kidney Disease.

No.	Variable	Significado	Valores	Unidad/Rango	Faltantes
1	age	Edad	Años	Numérica	9
2	bp	Presión arterial	mm/Hg	Numérica	12
3	sg	Gravedad específica	1.005 – 1.025	Nominal	47
4	al	Albumina	0, 1, 2, 3, 4, 5	Nominal	46
5	su	Azúcar	0, 1, 2, 3, 4, 5	Nominal	49
6	rbc	Glóbulos rojos	normal, anormal	Nominal	152
7	pc	Células de pus	normal, anormal	Nominal	65
8	pcc	Grupos de células pus	presente, no presente	Nominal	4
9	ba	Bacterias	presente, no presente	Nominal	4
10	bgr	Glucosa en sangre	mg/dL	Numérica	44
11	bu	Urea en la sangre	mg/dL	Numérica	19
12	sc	Suero de creatina	mg/dL	Numérica	17
13	sod	Sodio	mEq/L	Numérica	87
14	pot	Potasio	mEq/L	Numérica	88
15	hemo	Hemoglobina	gms	Numérica	52
16	pcv	Volumen glóbulos rojos	Pendiente	Numérica	71
17	wbcc	Núm. glóbulos blancos	Cell/cmm	Numérica	106
18	rbcc	Núm. glóbulos rojos	millones/cmm	Numérica	131
19	htn	Hipertensión	sí, no	Nominal	2
20	dm	Diabetes mellitus	sí, no	Nominal	2
21	cad	Arteriopatía coronaria	sí, no	Nominal	2
22	appet	Apetito	bueno, malo	Nominal	1
23	pe	Edema de piel	sí, no	Nominal	1
24	ane	Anemia	sí, no	Nominal	1
25	class	Clase (Target)	*ckd, notckd*	Nominal	0

4.2.2. Herramientas

Para desarrollar el análisis y modelado de los datos se utilizó la biblioteca Scikit-learn de Python. Se trata de una biblioteca de código abierto para el aprendizaje automático que permite tanto el aprendizaje supervisado como el no supervisado. Contiene clases y funciones para crear modelos, realizar preprocesamiento de datos, validación, entre otras (Pedregosa et al., 2011). En particular, se utilizó la biblioteca Pandas para manipular y analizar los datos, así como las bibliotecas Matplotlib y Seaborn para visualizar cada una de las variables.

4.2.3. Algoritmos clásicos de Aprendizaje automático

1. Árboles de decisión (*Decision tree*): Estos modelos predictivos se basan en reglas binarias de decisión (Sí/No), lo que permite clasificar las observaciones según sus atributos y estimar el valor de la variable objetivo. Los métodos basados en árboles son técnicas supervisadas no paramétricas que segmentan el espacio de los predictores en regiones más sencillas, lo que permite un análisis más claro de las interacciones entre las variables (Bagnato, 2020).

2. Naïve Bayes Gaussiano (*Gaussian Naïve Bayes*): Se trata de un modelo probabilístico, es decir, las predicciones se basan en calcular probabilidades. Näive Bayes Gaussiano asume que los datos siguen una distribución gaussiana. En el fondo, supone una distribución normal para cada variable y la clase, proporcionando el promedio y la desviación típica de los datos obtenidos.

3. *Gradient boosting*: Es un método de ensamble basado en la técnica de *boosting*. Es ampliamente utilizado para problemas de regresión y clasificación, y se considera uno de los métodos más robustos para construir modelos predictivos precisos. Este enfoque combina varios modelos para crear uno más eficiente para la predicción.

4. Máquinas de vector soporte (SVM, *Support Vector Machine*): Es un algoritmo supervisado de alto rendimiento con poca información. En este método se pretende encontrar un hiperplano que mejor separe las dos clases. Debido a que se puede tener un número infinito de hiperplanos que pasan por un punto, el SVM identifica el óptimo al encontrar el máximo margen, lo que significa establecer distancias máximas entre las dos clases (**Deitel2019**).

4.2.4. Algoritmos interpretables

1. *Explainable Boosting Classifier* (EBC): Representa una clase de modelos estadísticos que generalizan la regresión lineal al permitir que cada característica posea su propia función de contribución, lo que facilita la interpretación y la captura de relaciones no lineales de manera controlada. Es un modelo de aprendizaje automático que combina interpretabilidad y precisión, fundamentado en los Modelos Aditivos Generalizados (GAMs).

Un GAM es un tipo de modelo estadístico que generaliza el modelo lineal aditivo; a diferencia de los modelos lineales, los GAMs permiten capturar relaciones no lineales entre cada variable de entrada y la salida sin perder la interpretabilidad, expresando la predicción como una suma de funciones de cada característica . El EBC es una implementación particular de GAMs que utiliza *boosting* para ajustar las funciones $f_i(x_i)$ de cada característica. Este algoritmo aplica *boosting* en una versión simplificada, donde cada función f_i se ajusta gradualmente para mejorar la predicción del modelo de forma independiente para cada característica, permitiendo un enfoque altamente interpretable donde cada variable se considera por separado (Lakkaraju et al., 2017).

4.2.5. Medidas de rendimiento

Las medidas de rendimiento son métricas que se utilizan para evaluar la eficacia y exactitud de un modelo. Las métricas usadas en este trabajo son (Géron, 2022):

- *Accuracy*: Proporción de predicciones correctas sobre el total de predicciones.

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.1}$$

- *Precision*: Es utilizada para conocer qué porcentaje de valores que se han clasificado como positivos son realmente positivos.

$$\text{precision} = \frac{TP}{TP + FP} \tag{4.2}$$

- *Recall*: También conocida como el ratio de verdaderos positivos, es utilizada para conocer cuántos valores positivos son correctamente clasificados.

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

- **Error absoluto medio (MAE)**: Mide la magnitud media de los errores en un conjunto de predicciones, sin considerar su dirección. Mide la precisión de las variables continuas; es decir, el MAE es el promedio sobre la muestra de verificación de los valores absolutos de las

diferencias entre la predicción y la observación correspondiente. El MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio (Myers et al., 2012).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.4}$$

Donde n representa el número de observaciones, y_i representa el valor real y $\hat{y}_i$ representa el valor predicho.

4.2.6. Diseño experimental

En la Figura 4.1 se detalla la arquitectura del diseño experimental implementado para la clasificación de pacientes con ERC.

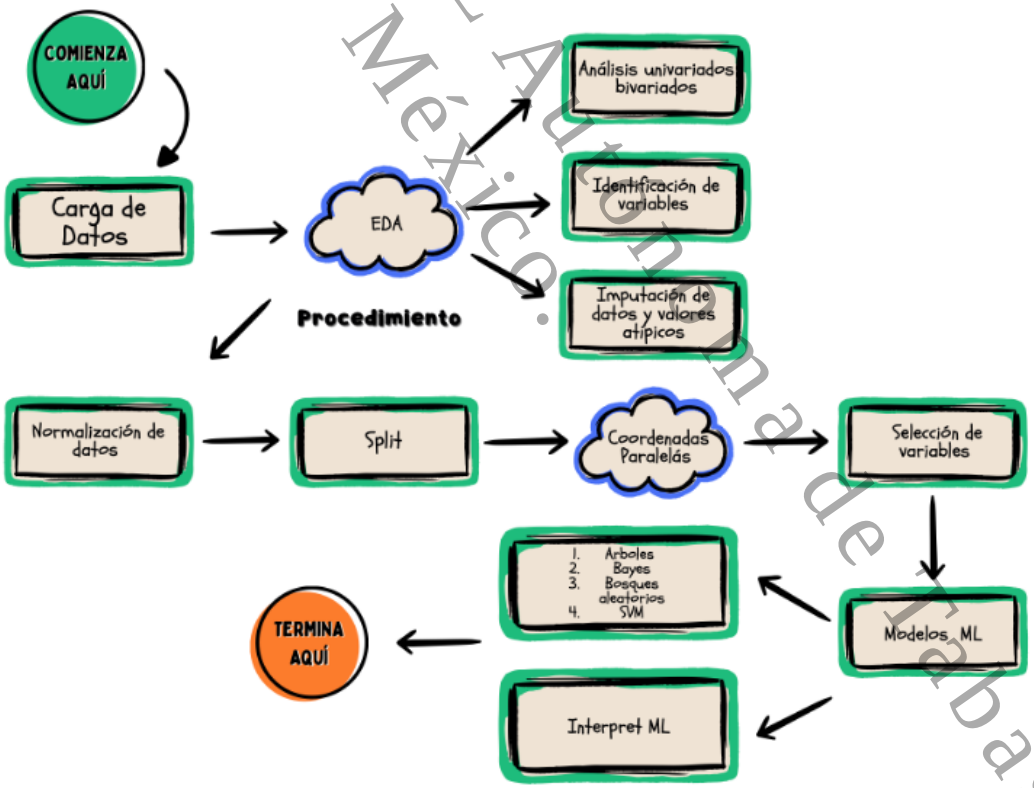

Figura 4.1. Flujo del diseño experimental propuesto para la clasificación de pacientes con ERC.

Como primer paso se describe el dataset que se utilizó para el modelado, usando los algoritmos descritos previamente y siguiendo el procedimiento estándar de *Scikit-Learn*:

- Separación de los datos en datos de prueba ($1/3$ de los datos) y datos de entrenamiento ($2/3$ de los datos).
- Imputación de los datos (limpieza de datos).
- Codificación de las variables nominales.

Posteriormente se evaluaron cuatro algoritmos de aprendizaje automático considerados clásicos en la literatura, junto con un algoritmo interpretable llamado *Explainable Boosting Classifier* (EBC), considerando los siguientes aspectos de diseño:

- Clasificador de Árbol de decisión:
  - Ideal para salidas categóricas.
  - No obstante, puede crear árboles complejos.
- Clasificador de Naïve Bayes ingenuo gaussiano:
  - Eficiente y rápido.
  - Puede ser menos confiable cuando hay dependencias entre los atributos.
- Gradient boosting:
  - Alta precisión predictiva.
  - Manejo de datos complejos.
- Clasificador SVM:
  - Capaz de modelar límites de decisión no lineales.
  - Bastante robusto contra el sobreajuste en un espacio de alta dimensión.
- Explainable Boosting Classifier (EBC):
  - Interpretabilidad transparente, analiza de forma global y local.
  - Uso intuitivo y herramientas visuales.

4.3. Interpretabilidad

La interpretabilidad es el grado en que un ser humano puede entender la causa de una decisión (Miller, 2017). Otra definición es que la interpretabilidad es el grado en que un ser humano puede predecir consistentemente el resultado del modelo (Kim et al., 2016). Se considera que un modelo es más fácil de entender que otro cuando sus decisiones son más comprensibles para una persona. A mayor interpretabilidad de un modelo de aprendizaje automático, será más sencillo entender las razones detrás de sus decisiones o predicciones.

La interpretabilidad se puede realizar de acuerdo con los criterios intrínsecos (por sí mismos). La interpretabilidad intrínseca se refiere a modelos de aprendizaje automático que son interpretables por sí mismos, o por la naturaleza de sus mismas estructuras, tales como los árboles de decisión y modelos lineales. La interpretabilidad intrínseca de un modelo se puede hacer de forma global o local (Lakkaraju et al., 2017). Un modelo de aprendizaje automático explicable es uno que cuenta con características o propiedades que facilitan la transparencia, la facilidad de comprensión y la capacidad de cuestionar o consultar los resultados del aprendizaje automático.

4.3.1. Interpretabilidad global

Se trata de entender cómo el modelo toma la decisión en forma general, considerando todas las variables, permitiendo comprender la relación de las entradas con las de la salida del modelo. Este nivel de interpretabilidad implica comprender cómo el modelo toma decisiones, basándose en una visión holística de sus características y cada uno de los componentes aprendidos, como pesos, otros parámetros y estructuras. Los modelos más simples, como la regresión lineal o los árboles de decisión, suelen ser más interpretables a nivel global porque sus estructuras y coeficientes son más fáciles de entender. La interpretabilidad del modelo global ayuda a comprender la distribución de su resultado objetivo en función de las características. La interpretabilidad del modelo global es muy difícil de lograr en la práctica (Gunning, 2017).

4.3.2. Interpretabilidad local

Se centra en comprender las predicciones del modelo para casos específicos o instancias individuales. Proporciona detalles sobre qué características específicas y cuánto influyeron en la

predicción para esa instancia particular. Es particularmente útil para modelos complejos como las redes neuronales y los métodos combinados (ensambles), donde la interpretación global puede ser difícil (Molnar, 2024).

4.3.3. Gráfica de coordenadas paralelas

Se trata de una técnica gráfica donde cada observación o punto de datos se ilustra como una línea que se desplaza por una serie de ejes paralelos, que corresponden a una variable de n dimensiones. Esto posibilita la indagación de vínculos, patrones y fluctuaciones que podrían aclararse en los datos sin procesar (Tufte, 2001). Los diagramas de coordenadas paralelas resultan beneficiosos en diversos contextos donde entender las relaciones multivariantes es esencial (Itoh et al., 2017):

- Exploración de datos multidimensionales: los gráficos de coordenadas paralelas son ideales para conjuntos de datos con múltiples variables donde es esencial comprender las interacciones y correlaciones entre estas variables.
- Análisis de características: en el análisis de datos y el aprendizaje automático, estos gráficos pueden ayudar a explorar cómo las diferentes características contribuyen a resultados específicos, lo cual es invaluable para la selección de características y la interpretación del modelo.
- Detección de anomalías: Los gráficos de coordenadas paralelas pueden resaltar anomalías y valores atípicos que pueden no ser evidentes en gráficos de variables individuales.

4.4. Trabajos relacionados

4.4.1. Trabajos que utilizan el mismo conjunto de datos

De acuerdo con la literatura revisada, se puede observar que hay trabajos relacionados con el mismo dataset usado en esta investigación, obteniendo en cada uno de ellos buenos resultados en su rendimiento, como se muestra en la Tabla 4.2. En esta Tabla se muestran los 10 trabajos más relevantes de un total de 24 artículos analizados.

Tabla 4.2. Trabajos previos que utilizaron el conjunto de datos CKD del repositorio UCI.

Autor	**Algoritmos**	*Accuracy*	**Interpretabilidad**
Rubini et al., 2015	Radial Basis Function network MLP (Red neuronal profunda) Regresión logística	98.50 99.75 97.50	Parcial (Regresión logística)
Avci et al., 2018	Naïve Bayes K-Star SVM J48	95.00 91.75 97.75 99.00	Parcial (J48, Naïve Bayes)
Qin et al., 2020	Regresión logística Random Forest SVM k-NN (k-Vecinos más cercanos) Naïve Bayes Feed Forward Neural Network	97.00 99.75 99.25 99.25 95.75 99.50	Parcial (Naïve Bayes, Regresión logística)
P. Ghosh et al., 2020	SVM Ada Boost Linear Discriminate Analysis Gradient boosting	99.56 97.91 97.91 99.80	Parcial (Linear Discriminate Analysis)
Shankar et al., 2020	Wide & Deep Learning Regresión logística k-NN Feed Forward Neural Network Random forest	98.00 96.00 97.00 99.00 99.00	Parcial (Regresión logística)
Almustafa, 2021	k-NN Naïve Bayes J48 Random forest Decision table SGD	95.75 95.00 99.00 95.50 99.00 98.25	Parcial (J48, Naïve Bayes)
Singh et al., 2022	MLP	100	Ninguna
Gokiladevi et al., 2022	SVM Random forest Regresión logística k-NN Árboles de decisión	85.00 99.00 97.00 79.00 96.00	Parcial (Árboles de decisión)
Pal, 2023	Árboles de decisión SVM	95.92 97.23	Parcial (Árboles de decisión)
Kaur et al., 2023	Random forest Árboles de decisión k-NN SVM MLP Regresión logística Naïve Bayes	98.33 93.89 95.12 97.45 96.78 94.56 92.34	Parcial (Árboles de decisión, Naïve Bayes)

El primer trabajo realizado con el CKD dataset fue en el año 2015 y fue realizado por los creadores del conjunto de datos, quienes aparecen en el primer lugar de nuestra tabla Rubini et al., 2015 abordan la predicción temprana de la enfermedad renal crónica (CKD, por sus siglas en inglés), donde compararon tres algoritmos de clasificación: Radial Basis Function networks, Multi-Layer Perceptron (MLP) y Regresión logística, evaluando su precisión. Se obtuvo una mejor precisión de 99.75 % en MLP, seguido de RBFN con 98.50 % y Regresión logística con 97.50 %. Se concluye en este trabajo que MLP es el mejor clasificador para la predicción temprana de CKD. Sin embargo, dentro de este artículo no se menciona ni se usa ningún algoritmo explicativo para la interpretación de los modelos, centrándose únicamente en el rendimiento de los clasificadores.

Por otro lado, Avci et al., 2018 usaron el software WEKA para la clasificación de los datos. Se utilizaron los clasificadores Naïve Bayes, K-star, SVM y J48. Se concluye que de los clasificadores usados, el más eficiente fue J48 con un rendimiento de 99.00 %. El objetivo de este trabajo es comparar el rendimiento en términos de precisión y eficiencia, por lo que carece de interpretabilidad de los modelos.

En Qin et al., 2020 utilizaron seis algoritmos: Regresión logística, Random Forest, SVM, k-Nearest Neighbors (k-NN), Naïve Bayes y Feed Forward Neural Network para obtener el mejor rendimiento, el cual fue Random Forest con una precisión de 99.75 %. Además, se propuso un modelo integrado que combina regresión logística y Random Forest utilizando un perceptrón, logrando una precisión promedio de 99.83 % después de 10 simulaciones. Ninguno de los algoritmos utilizados se considera explicativo.

Por su parte P. Ghosh et al., 2020 implementaron cuatro algoritmos de aprendizaje automático: SVM, AdaBoost (AB), Linear Discriminant Analysis (LDA) y Gradient boosting (GB). Se evaluaron los modelos resultantes utilizando métricas de rendimiento como precisión, sensibilidad y especificidad. El algoritmo GB obtuvo la mayor precisión con un 99.80 %. Los otros algoritmos también mostraron buenos resultados, pero GB se destacó en términos de precisión y rendimiento general. Sin embargo, el modelo resultante se considera de caja negra.

Shankar et al., 2020 crearon diversos modelos de aprendizaje automático para identificar patrones y factores de riesgo asociados con la enfermedad. Implementaron técnicas de preprocesamiento de datos para manejar valores faltantes y normalizar las variables, garantizando así la calidad y consistencia del conjunto de datos. Los modelos de aprendizaje automático demostraron

ser herramientas efectivas para la predicción de la ERC, permitiendo una identificación temprana de la enfermedad. Sin embargo, estos modelos carecen de interpretabilidad.

Almustafa, 2021 evaluaron varios algoritmos de clasificación, incluyendo Árboles de decisión, Decision table, k-NN, J48, Stochastic Gradient Descent (SGD) y Naïve Bayes. Además, aplicaron un modelo de predicción basado en la selección de características para mejorar la eficiencia en la predicción de casos de CKD. El modelo de predicción propuesto mostró una alta precisión en la detección temprana de CKD, lo que puede ayudar a prevenir el empeoramiento de la enfermedad. Los resultados indicaron que la selección de características y el uso de algoritmos de clasificación adecuados pueden mejorar significativamente la precisión de las predicciones. Se concluye en este trabajo que los algoritmos J48 y Decision table fueron los mejores clasificadores para la predicción temprana de CKD. En este caso, los algoritmos utilizados pueden ser considerados parcialmente explicativos.

Singh et al., 2022 desarrollaron un modelo de red neuronal profunda para la detección temprana y predicción de la enfermedad renal crónica (CKD). Las características más importantes fueron seleccionadas mediante la eliminación recursiva de características (RFE). Las características clave incluyeron hemoglobina, gravedad específica, creatinina sérica, conteo de glóbulos rojos, albúmina, volumen celular empaquetado e hipertensión. El modelo de red neuronal profunda propuesto superó a otros cuatro clasificadores (SVM, k -NN, Regresión logística, Random forest y Naïve Bayes) al lograr una precisión del 100 %. Como mencionamos anteriormente, los modelos basados en redes neuronales son los modelos menos explicativos que existen.

Gokiladevi et al., 2022 evaluaron cinco algoritmos de clasificación de aprendizaje automático: SVM, Random Forest, Regresión logística, k-NN y Árboles de decisión. El preprocesamiento de datos incluyó la sustitución de valores faltantes y la transformación de datos. Los modelos fueron evaluados utilizando métricas de rendimiento como precision, recall y F-score. Como en otros trabajos, los mejores modelos resultaron ser de caja negra.

Pal, 2023 aplicó tres clasificadores de aprendizaje automático: Regresión logística, Árboles de decisión y SVM. Además, se utilizó el método de ensamble de *bagging* para mejorar los resultados del modelo desarrollado. El clasificador de Árboles de decisión obtuvo una precisión del 95.92 %. Después de aplicar el método de ensamble de bagging, la precisión aumentó a 97.23 %, lo que indica una mejora significativa en la capacidad predictiva del modelo. El detalle al aplicar *bagging*

es que el modelo pierde interpretabilidad.

Un último trabajo relevante fue realizado en el año 2023, que está ubicado en el último lugar de nuestra Tabla y demuestra la importancia y vigencia del dataset. Kaur et al., 2023 utilizaron varios algoritmos de aprendizaje automático para predecir la enfermedad: k-NN, Regresión logística, Árboles de decisión, Random forest, Naïve Bayes, SVM y MLP. Los datos fueron preprocesados para manejar valores faltantes y normalizar las características. Se aplicaron técnicas de validación cruzada para evaluar el rendimiento de los modelos. El modelo de Random forest obtuvo la mayor precisión en la predicción de CKD, seguido de cerca por SVM y MLP. Los modelos resultantes carecen de interpretabilidad.

Observando los resultados de los trabajos que utilizan el mismo conjunto de datos, el que obtuvo mejor precisión fue hecho por Singh et al., 2022, con una red neuronal, seguido por el trabajo de P. Ghosh et al., 2020, ambos con algoritmos tradicionales de Aprendizaje automático. La mayoría de los trabajos analizados tienen interpretabilidad parcial, ya que los algoritmos utilizados (Árboles de decisión, Naïve Bayes y Regresión logística), son interpretables por naturaleza, aunque realmente no se menciona nada de interpretabilidad de los modelos en ninguno de los trabajos previos a este trabajo. De igual manera, los mejores modelos obtenidos en todos los trabajos resultaron ser de caja negra, es decir, no interpretables.

4.4.2. Trabajos que utilizan otros conjuntos de datos

Existen ciertos trabajos para la predicción de la ERC utilizando otros conjuntos de datos. Por ejemplo, Debal y Sitote, 2022 se enfocaron en la detección temprana y precisa de las etapas de la ERC para minimizar las complicaciones de salud de los pacientes. Los algoritmos utilizados incluyeron Random forest, SVM y Árboles de decisión. Se aplicaron técnicas de selección de características y validación cruzada para evaluar el rendimiento de los modelos. Los resultados indicaron que el modelo Random forest obtuvo un mejor desempeño en comparación con los otros 2 algoritmos.

Por su parte, Bai et al., 2022 evaluaron la viabilidad de utilizar técnicas de Aprendizaje automático para predecir el riesgo de enfermedad renal en etapa terminal en pacientes con ERC. Se utilizaron datos de una cohorte longitudinal de ERC y se entrenaron cinco algoritmos: Regre-

sión logística, Naïve Bayes, Random Forest, Árboles de decisión y k-Nearest Neighbors (k-NN). Los modelos se compararon contra la Kidney Failure Risk Equation (KFRE). Los resultados mostraron que tres modelos resultantes tenían una predictibilidad equivalente y mayor sensibilidad en comparación con la KFRE.

4.5. Resultados y discusión

La Tabla 4.3 muestra los resultados de los experimentos realizados con los cuatro algoritmos clásicos de aprendizaje automático y con el algoritmo explicativo seleccionado.

Tabla 4.3. Resultados de los experimentos.

Clasificador	Accuracy	Precision	Recall
Naïve Bayes Gaussiano	0.98	1.00	0.96
SVM	0.98	0.96	0.97
Árboles de decisión	0.99	0.99	0.97
Gradient boosting	0.99	1.00	0.99
Explainable Boosting Classifier (EBC)	0.99	1.00	0.99

Los resultados obtenidos indican que los modelos tradicionales de aprendizaje automático: SVM, Naïve Bayes, Árboles de decisión y Gradient boosting tienen un desempeño sólido en la clasificación de pacientes con enfermedad renal crónica del riñón (ERC), con valores de precisión de al menos 0.98. Sin embargo, estos modelos varían en términos de interpretabilidad y aplicabilidad en el contexto médico. En particular, el modelo Explainable Boosting Machine (EBM) logra un rendimiento comparable a los mejores modelos tradicionales (0.99 de precisión), pero con la ventaja adicional de ofrecer interpretabilidad tanto local como global en sus predicciones. A diferencia de modelos como SVM y Gradient boosting, que funcionan como cajas negras, EBM permite identificar de manera transparente qué factores influyen en la predicción de la enfermedad, lo que facilita su adopción en entornos clínicos donde la toma de las decisiones es crucial. Aunque Gradient boosting y Árboles de decisión también obtuvieron altos valores de *precision* y *recall*, su nivel de explicabilidad es menor en comparación con EBM. Por otro lado, aunque Naïve Bayes mostró un desempeño competitivo, su simplicidad podría limitar su capacidad para capturar relaciones más complejas en los datos. En conclusión, el uso de EBM en el análisis de ERC permite mantener una alta precisión sin comprometer la interpretabilidad o dando una inter-

pretabilidad tanto local como global, lo que lo convierte en una alternativa viable para mejorar la transparencia en el diagnóstico basado en IA.

4.5.1. Gráficas de coordenadas paralelas

Con el propósito de evaluar visualmente la importancia de los atributos, se generó la gráfica de coordenadas paralelas (Figura 4.2), la cual permite identificar de manera intuitiva qué variables presentan un mayor poder predictivo en el modelo. Esta representación facilita la detección de patrones y diferencias significativas entre las clases *ckd* y *notckd*.

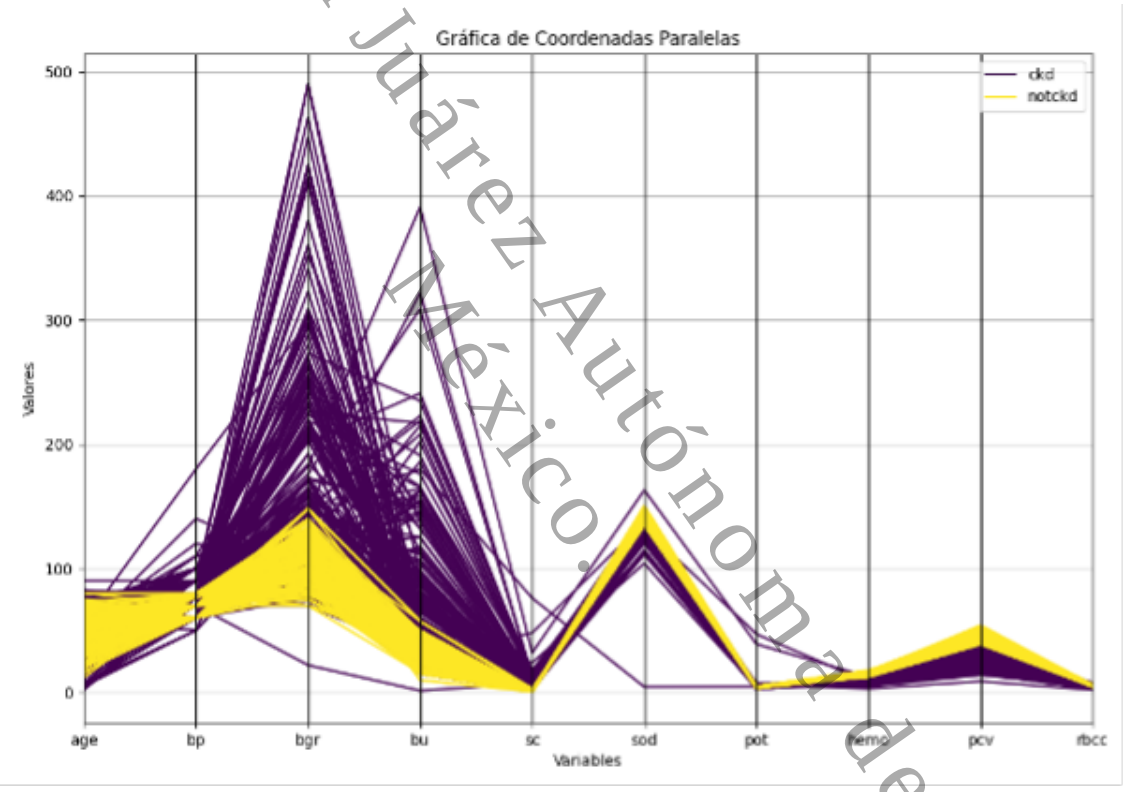

Figura 4.2. Gráfica de coordenadas paralelas para la visualización de patrones multivariantes en pacientes con y sin ERC.

Para este caso, no se normalizaron ni estandarizaron los valores de las variables. Esto significa que los valores reales están representados en su escala original, lo que permite interpretar directamente los rangos y las diferencias entre las clases. La gráfica se puede interpretar de la siguiente manera:

- `bgr` (glucosa en la sangre): Se observa una clara diferencia entre las clases; la mayoría de

las instancias de la clase *notckd* (amarillo) tienen valores más bajos de glucosa, mientras que las de la clase *ckd* (morado) tienden a tener valores más altos . Esto sugiere que `bgr` es una variable relevante para la predicción.

- `sod` (sodio): Muestra una separación moderada entre las clases *notckd* y *ckd*. Las instancias de *notckd* presentan valores más consistentes y elevados en comparación con las de *ckd*.
- `pot` (potasio): Las líneas de ambas clases están bastante mezcladas, lo que sugiere que `pot` no es una variable con alto poder predictor.
- `hemo` (hemoglobina): Se aprecia una separación clara entre las clases. Las líneas amarillas (*notckd*) se concentran en valores más altos, mientras que las moradas (*ckd*) se sitúan en valores más bajos, resaltando su importancia como una predictora fuerte.
- `bu` (urea en la sangre): Las instancias de *notckd* parecen tener valores más bajos que las de *ckd*, aunque existe mezcla entre ellas; por lo tanto, aporta menos a la predicción.
- `age` (edad) y `bp` (presión arterial): En estas variables, las líneas de ambas clases se mezclan significativamente, lo que sugiere una menor capacidad predictora.

En general, podemos concluir que:

- Variables como `bgr`, `hemo` y `sod` muestran diferencias más claras entre las clases, lo cual implica que son características potenciales para el modelo.
- Variables como `pot`, `age` y `bp` tienen patrones menos diferenciados, indicando que probablemente contribuyen menos al poder predictivo global.

4.5.2. Análisis de correlación

Para identificar las variables más influyentes en el diagnóstico de insuficiencia renal, se llevó a cabo un análisis de correlación entre las características del conjunto de datos y la variable objetivo (`class`). La Figura 4.3 muestra el mapa de calor de correlación, donde se observa la relación entre las variables. En particular, se destaca la correlación de ciertas características con la variable objetivo o clase, la cual representa la presencia o ausencia de insuficiencia renal.

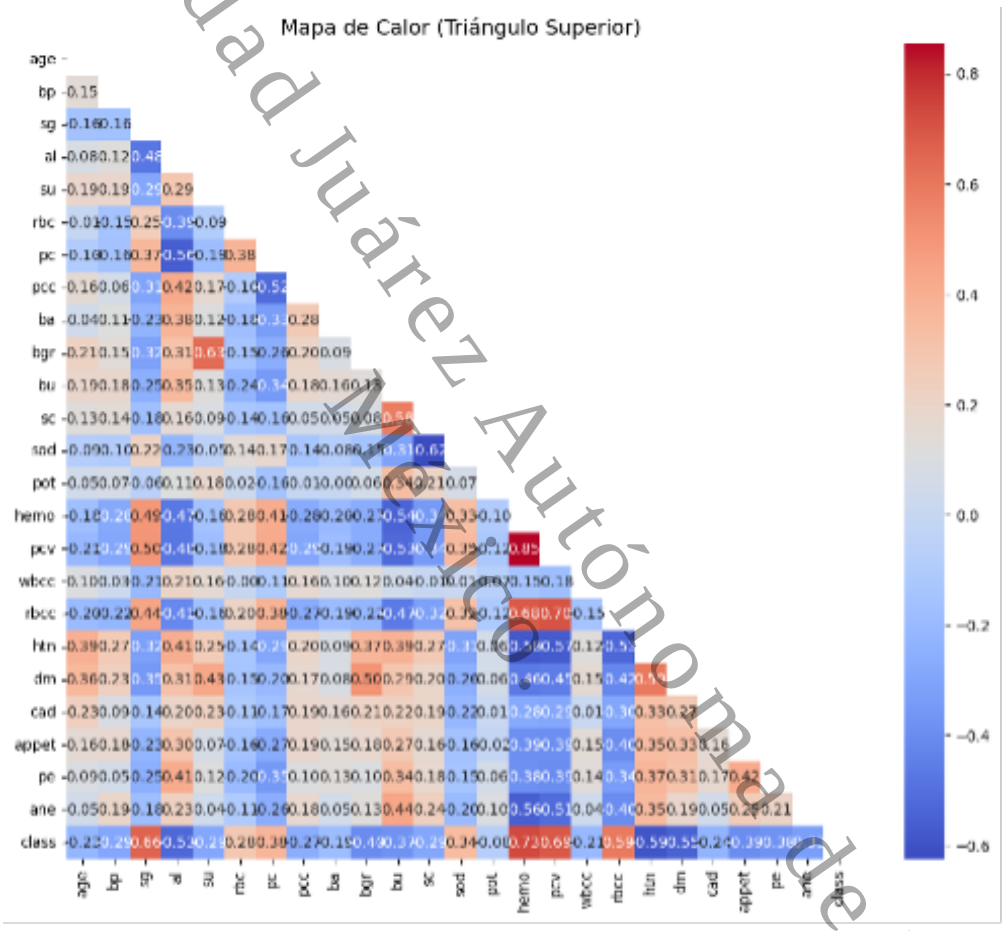

Figura 4.3. Mapa de calor con la correlación entre las variables con ERC.

En la Tabla 4.4 se describen los valores relevantes de la correlación, Estas variables presentan correlaciones moderadas a altas con la insuficiencia renal, lo que sugiere que pueden desempeñar un papel clave en la predicción de la enfermedad.

Tabla 4.4. Resultados de correlación.

Variable	Coeficiente de correlación
Hemoglobina (`hemo`)	0.73
Volumen de células compactas (`pcv`)	0.69
Gravedad específica de la orina (`sg`)	0.65
Conteo de glóbulos rojos (`rbcc`)	0.59
Hipertensión (`htn`)	0.59
Diabetes mellitus (`dm`)	0.54
Albúmina (`al`)	0.53

De estos resultados podemos decir lo siguiente, analizando cómo cada variable se correlaciona con la clase:

- Hemoglobina (`hemo`) y `pcv` (volumen de células compactas): La alta correlación positiva indica que niveles bajos de hemoglobina y hematocrito están asociados con la insuficiencia renal. Esto concuerda con la literatura médica, donde se ha documentado que la anemia es una complicación frecuente en pacientes con insuficiencia renal crónica.
- Gravedad específica de la orina (`sg`): Su alta correlación sugiere que las alteraciones en la concentración de la orina pueden ser un indicador temprano de insuficiencia renal.
- Hipertensión (`htn`) y Diabetes Mellitus (`dm`): Ambas enfermedades son factores de riesgo bien conocidos para el desarrollo de insuficiencia renal, lo que se refleja en su fuerte asociación con la variable objetivo.
- Baja correlación de otras variables: Algunas características, como *pot* (potasio), presentan coeficientes de correlación bajos, lo que sugiere que su impacto en la insuficiencia renal es menor en este conjunto de datos.

Los resultados obtenidos indican que ciertos parámetros clínicos y de laboratorio pueden ser utilizados para mejorar la predicción de insuficiencia renal en modelos de aprendizaje automático. En particular, variables como `hemo`, `pcv`, `sg`, `rbcc`, `htn` y `dm` deberían considerarse como factores clave en la construcción de modelos predictivos.

4.5.3. Análisis de interpretabilidad global de EBM

Como contribución a los trabajos realizados en investigaciones anteriores, este estudio proporciona una explicación de la predicción del modelo utilizando el algoritmo EBM, tanto a nivel global como local. En la Figura 4.4 se presentan las características de todo el conjunto de datos, donde se visualiza cuál variable es la más predictiva para el modelo.

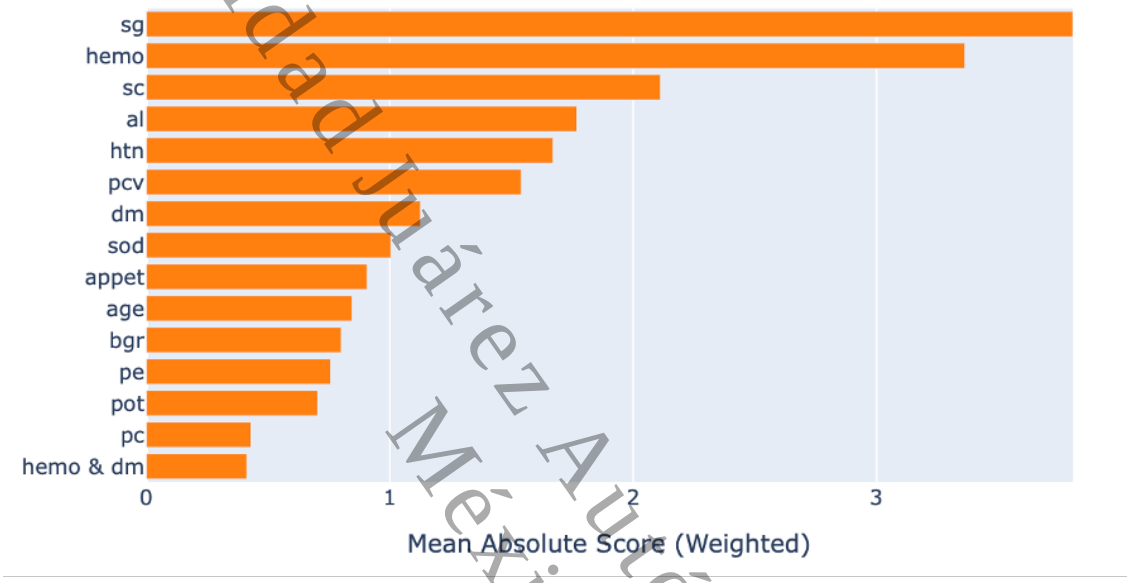

Figura 4.4. Gráfica de interpretabilidad global.

La gráfica de la Figura 4.4 refleja la importancia promedio de cada característica en el modelo. Se calcula utilizando la *Mean Absolute Score (Weighted)*, la cual indica cuánto influye cada característica en las predicciones, considerando todos los casos del conjunto de datos.

En el eje vertical se presentan las características ordenadas de mayor a menor importancia para la predicción del modelo. En el eje horizontal se muestra el promedio de la contribución de cada característica a las predicciones, tomando en cuenta todo el conjunto de datos. Los valores más altos, como `sg`, `hemo` y `sc`, indican que estas características tienen un fuerte impacto y son indicadores clave en la detección de la enfermedad renal crónica. Las variables con valores medios, como `al`, `htn` y `pcv`, muestran una influencia moderada en la detección de la enfermedad. Por otro lado, las características con menor impacto, o que aportan poco al modelo global, son `pot` y `pe`.

4.5.4. Análisis de interpretabilidad local de EBM

En esta sección se analiza cada característica individualmente para determinar cuál tiene mayor influencia en la predicción de un paciente en particular. Para ilustrar este concepto, se examinan casos donde el modelo acierta y donde presenta errores.

Caso 1: Predicción correcta de enfermedad (*True Positive*)

A continuación, analizamos un paciente específico donde el modelo predijo correctamente la clase *ckd* con una probabilidad del 100 % (Figura 4.5) En la gráfica se observa que tanto la clase real como la predicción coinciden en *ckd* con total confianza.

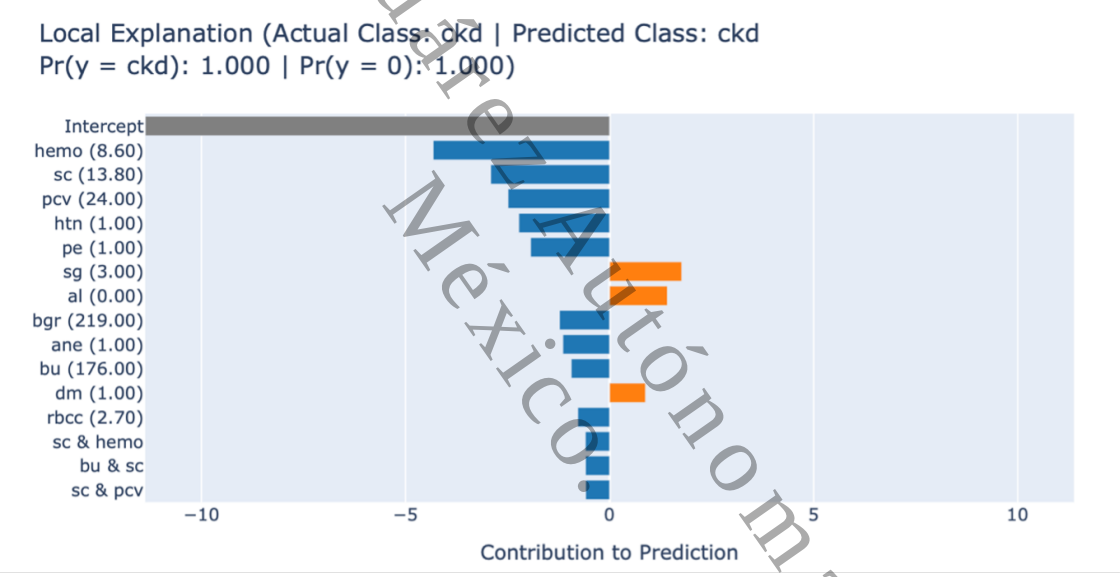

Figura 4.5. Gráfica de interpretabilidad local donde el modelo acierta en la predicción *ckd*.

En el eje horizontal se representan las contribuciones de cada característica; las barras azules tienen un impacto negativo que favorece la predicción de *ckd* (paciente enfermo), mientras que las barras naranjas tienen un impacto positivo hacia *notckd* (paciente sano). El eje vertical presenta las características junto con su valor específico y el intercepto del modelo.

Las características con mayor impacto negativo, que refuerzan la predicción de *ckd*, son: `hemo`, `sc`, `pcv`, `bgr` y `bu`. Por otro lado, aunque `sg`, `al` y `dm` tienen un impacto positivo, su efecto es mínimo, por lo que la suma total de las contribuciones confirma el diagnóstico de insuficiencia renal en este paciente.

Caso 2: Predicción correcta de salud (*True Negative*)

Ahora analizaremos un caso en el que el modelo predice la clase *notckd* con una probabilidad del 100 % (Figura 4.6). En la parte superior de la gráfica se observa que el valor real es *notckd* y la predicción del modelo coincide plenamente con dicho estado.

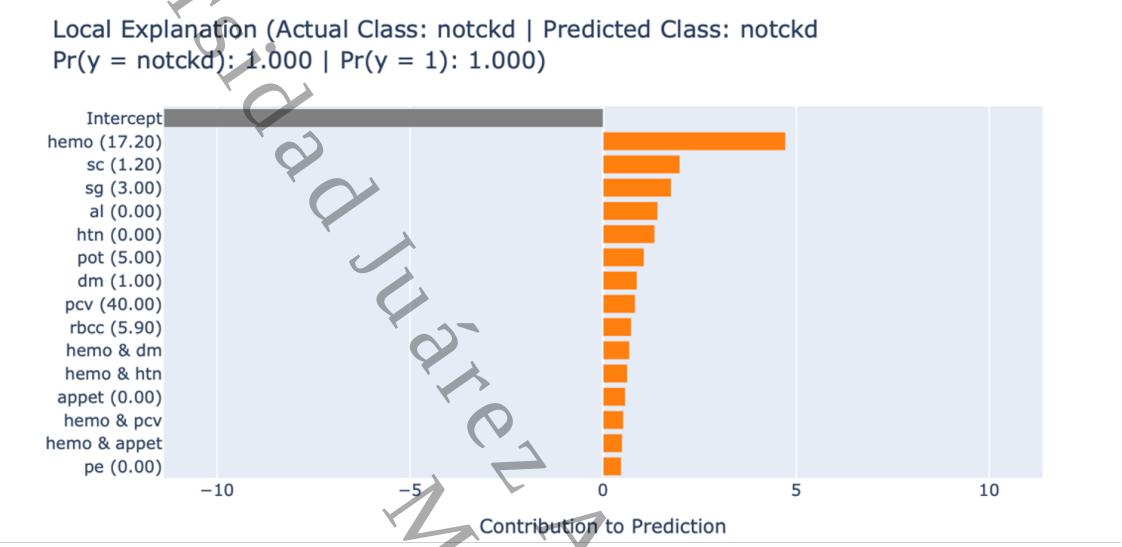

Figura 4.6. Gráfica de interpretabilidad local donde el modelo acierta en la predicción *notckd*.

En el gráfico, el eje horizontal representa las contribuciones de cada variable. Dado que todas las barras son de color naranja, esto indica que sus contribuciones son positivas, lo que lleva al modelo a predecir una condición sin enfermedad renal crónica. En el eje vertical se encuentran las variables predictoras de este paciente, organizadas de mayor a menor importancia. La característica con mayor impacto es `hemo`, mientras que la de menor influencia individual es `rbcc`. La predicción final es el resultado de la suma de todas las contribuciones positivas y el intercepto, lo que conlleva a una predicción con total confianza hacia la clase *notckd*. Esto indica que todas las características relevantes favorecen la clasificación del paciente como sano.

Caso 3: Error de predicción (*False Negative*)

A continuación se analiza el único caso en el que el modelo se equivoca al predecir que un paciente está sano cuando en realidad presenta enfermedad renal crónica. De hecho, esta es la única instancia en la que el modelo comete un error en todo el conjunto de datos (Figura 4.7). En este escenario, el modelo predice con un 56.6 % de probabilidad que el paciente no padece la

enfermedad (*notckd*), aunque con una confianza moderada.

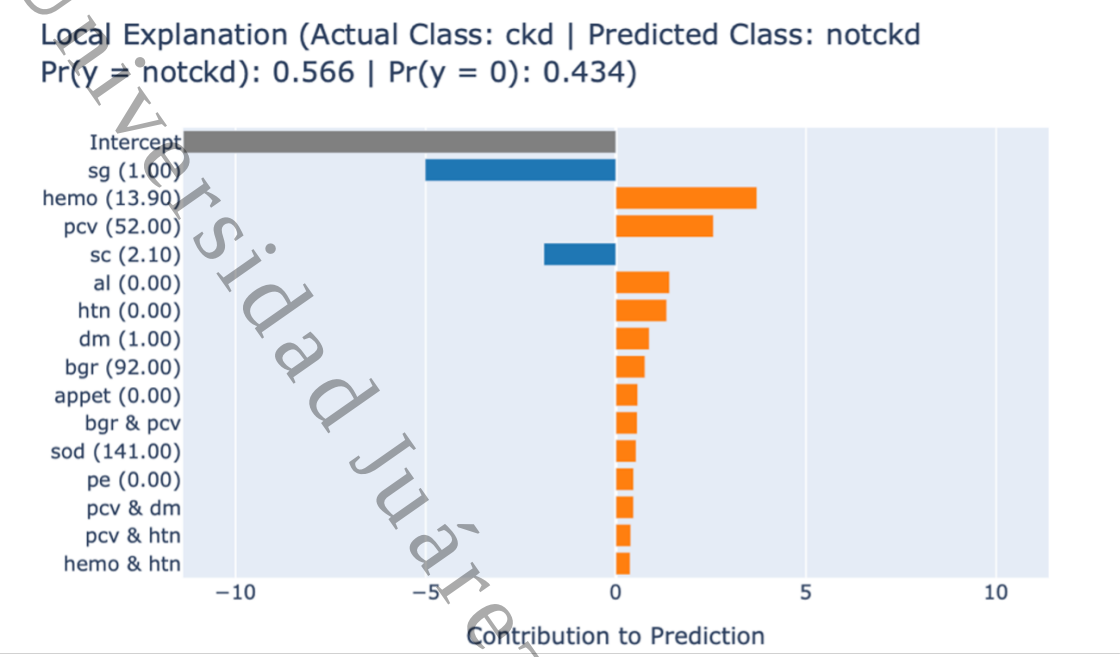

Figura 4.7. Gráfica de interpretabilidad local donde el modelo falla en la predicción.

Según el análisis de correlación, la variable hemo tiene mayor influencia que `sg`, lo que explica por qué el modelo clasifica al paciente como sano, cuando en realidad sí padece la enfermedad (*ckd*). Al analizar el gráfico, en el eje horizontal se observa cómo cada característica impacta en la predicción final. Las barras de color naranja contribuyen positivamente a la predicción de *notckd*, mientras que las barras de color azul influyen negativamente, favoreciendo la predicción de *ckd*. En el eje vertical se presentan las características más relevantes para este caso, junto con sus valores específicos. Se observa que `sg` (1.00) induce al modelo a predecir que el paciente está enfermo, pero como hay más características que favorecen la predicción de *notckd* (como `hemo` y `pcv`, entre otras), el resultado final es que el modelo clasifica al paciente como sano. Dado que la mayoría de las características impulsan positivamente la predicción de *notckd*, el modelo concluye que el paciente no tiene la enfermedad. Sin embargo, si consideramos la interpretación del modelo global, donde `sg` es la variable más predictiva, tendría sentido que el modelo hubiera producido un resultado *ckd* en lugar de *notckd*.

Discusión sobre la relevancia clínica del error

Desde el punto de vista médico, el error presentado en el caso 3 constituye un falso negativo, el escenario de mayor riesgo en el diagnóstico clínico. Mientras que otros errores pueden derivar en pruebas adicionales, la clasificación errónea de un paciente con ERC como sano (*notckd*) impide la intervención temprana, factor crucial para frenar la progresión de la enfermedad hacia estadios letales.

Este caso subraya la necesidad de la interpretabilidad local: al observar que la variable `sg` (gravedad específica) indicaba enfermedad pero fue superada por el peso de la hemoglobina (`hemo`), el especialista puede cuestionar la predicción de la IA y realizar una validación clínica manual, evitando las consecuencias potencialmente fatales de un diagnóstico omitido.

4.6. Conclusiones

Este trabajo demuestra la importancia del análisis exploratorio de datos y el uso de modelos de Aprendizaje automático explicables para predecir la insuficiencia renal crónica. En primer lugar, la gráfica de coordenadas paralelas proporciona un vistazo preliminar de las posibles correlaciones entre las variables y la clase. En segunda, a través de la implementación del algoritmo Explainable Boosting Machine (EBM), se logró combinar rendimiento y transparencia, ya que facilita tanto la interpretación global como local de las predicciones realizadas por el modelo (al contrario de los modelos tradicionales de aprendizaje automático, que no muestran qué variables son altamente predictoras). En este caso particular, esto implica poder informarle a un paciente cuáles son los elementos que más le afectan a su diagnóstico.

Las variables más influyentes en las predicciones, como `sg` (gravedad específica), `hemo` (hemoglobina) y `sc` (creatinina sérica), destacan como indicadores clave para identificar pacientes en riesgo, lo que valida su relevancia clínica en el diagnóstico de insuficiencia renal crónica. Por otro lado, características como `pot` (potasio) y `pe` (edema) tuvieron un impacto menor, lo que sugiere su limitada capacidad predictiva.

La capacidad de EBM de interpretar tanto las predicciones globales como locales del modelo permite no solo una mayor precisión en el diagnóstico, sino también la posibilidad de intervenir

de manera temprana en pacientes con alto riesgo. Esto no solo mejora los resultados clínicos, sino que también puede contribuir a la reducción de costos asociados al tratamiento de etapas avanzadas de la enfermedad.

Finalmente, los modelos explicativos no solo mejoran la precisión de los modelos predictivos, sino que también proporcionan una herramienta interpretativa útil para médicos y especialistas, promoviendo su uso en la práctica clínica y apoyando la toma de decisiones informada. Como trabajo futuro se considera recopilar un conjunto de datos propio a fin de ampliar las capacidades predictivas e interpretativas en problemas locales.

Referencias del capítulo

Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, *24*, 100631. https://doi.org/10.1016/j.imu.2021.100631

Arias Rodríguez, M., Aljama García, P., Egido de los Ríos, J., Lamas Peláez, S., Praga Terente, M., & Serón, D. (2013). *Nefrología clínica*. Editorial Médica Panamericana.

Avci, E., Karakus, S., Ozmen, O., & Avci, D. (2018). Performance comparison of some classifiers on Chronic Kidney Disease data. *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, 1-4. https://doi.org/10.1109/ISDFS.2018.8355392

Bagnato, J. I. (2020). Aprende Machine Learning en Español: Teoría + Práctica Python. https://www.aprendemachinelearning.com

Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific Reports*, *12*, 8377. https://doi.org/10.1038/s41598-022-12316-z

Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, *9*, 109. https://doi.org/10.1186/s40537-022-00657-5

El Heraldo de México. (2022). Supera Tabasco en 4 años más de mil muertes renales. https://oem.com.mx/elheraldodetabasco/ciencia-y-salud/supera-tabasco-en-4-anos-mas-de-mil-muertes-renales-19406978

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3.ª ed.). O'Reilly Media.

Ghosh, P., Shamrat, F. M. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020). Optimization of prediction method of chronic kidney disease using machine learning algorithm. *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1-6. https://doi.org/10.1109/iSAI-NLP51646.2020.9376787

Gokiladevi, M., Santhoshkumar, S., & Varadarajan, V. (2022). Machine learning algorithm selection for chronic kidney disease diagnosis and classification. *Malaysian Journal of Computer Science*, 102-115. https://doi.org/10.22452/mjcs.sp2022no1.8

Gunning, D. (2017). *Explainable artificial intelligence (xAI)* (inf. téc.). Defense Advanced Research Projects Agency (DARPA).

Instituto Nacional de Salud Pública. (2020). Enfermedad renal crónica en México. https://www.insp.mx/avisos/5296-enfermedad-renal-cronica-mexico.html

Itoh, T., Kumar, A., Klein, K., & Kim, J. (2017). High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *Journal of Visual Languages & Computing*, *43*, 1-13. https://doi.org/10.1016/j.jvlc.2017.03.001

Kaur, C., Kumar, M. S., Anjum, A., Binda, M. B., Mallu, M. R., & Al Ansari, M. S. (2023). Chronic Kidney Disease Prediction Using Machine Learning. *Journal of Advances in Information Technology*, *14*(2), 384-391. https://www.jait.us/uploadfile/2023/JAIT-V14N2-384.pdf

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*. https://doi.org/10.48550/arXiv.1707.01154

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269*. https://arxiv.org/abs/1706.07269

Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book

Myers, R. H., Walpole, R. E., & Myers, S. L. (2012). *Probabilidad y Estadística para ingeniería y ciencias* (9.ª ed.). PEARSON EDUCACIÓN.

Organización Panamericana de la Salud. (2021). Carga de enfermedades renales. https://www.paho.org/es/enlace/carga-enfermedes-renales

Pal, S. (2023). Chronic Kidney Disease Prediction Using Machine Learning Techniques. *Biomedical Materials & Devices*, *1*, 534-540. https://doi.org/10.1007/s44174-022-00027-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2020). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991-21002. https://doi.org/10.1109/ACCESS.2019.2963053

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Dataset. https://doi.org/10.24432/C5G020

Shankar, S., Verma, S., Elavarthy, S., Kiran, T., & Ghuli, P. (2020). Analysis and prediction of chronic kidney disease. *International Research Journal of Engineering and Technology*, *7*(5), 2395-0056. https://www.irjet.net/archives/V7/i5/IRJET-V7I5868.pdf

Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. *Diagnostics*, *12*(1), 116. https://doi.org/10.3390/diagnostics12010116

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd). Graphics Press.

World Kidney Day. (2022). Campaña de prevención renal 2022. https://www.worldkidneyday.org/events/campana-de-prevencion-renal-2022/

Capítulo 5

Detección temprana de nefropatía diabética empleando el algoritmo LIME

Detección temprana de nefropatía diabética empleando un algoritmo de Inteligencia Artificial interpretable

Simón Javier Hernández-Gaspar; Betania Hernández-Ocaña; José Hernández-Torruco; Oscar Chávez-Bosquez

Resumen El presente estudio presenta un modelo de Inteligencia Artificial para la predicción temprana de nefropatía diabética, una complicación renal grave de la diabetes tipo 2, utilizando el dataset DiabetIA del Instituto Mexicano del Seguro Social (IMSS) con 42 182 instancias. Durante el preprocesamiento realizado al dataset se seleccionaron 13 variables relevantes de un total de 520, se eliminaron variables con datos faltantes arriba del 40 %, se realizó imputación de datos y balanceo de clases mediante submuestro. Posteriormente, se compararon 5 algoritmos de aprendizaje automático (Árboles de decisión, SVM, Naïve Bayes, Gradient boosting y Random forest), siendo Random forest el que obtuvo el mejor rendimiento global con un rendimiento del 80 %. Para abordar el problema de la interpretabilidad, se aplicó la técnica LIME (*Local Interpretable Model-agnostic Explanations*) al modelo resultante con Random forest. El análisis con LIME demostró que las variables más influyentes en las predicciones fueron la hipertensión esencial y la edad del paciente. Además, se describe a detalle la interpretabilidad local de LIME en los 4 casos posibles: verdadero positivo, verdadero negativo, falso positivo y falso negativo. Se concluye que LIME es una herramienta útil para aumentar la confianza y transparencia en los sistemas de apoyo clínico, aunque se recomienda complementar el modelo con validación clínica para reducir errores en el diagnóstico.

Palabras clave: Aprendizaje supervisado, clasificación, interpretabilidad, LIME, Random forest.

Institución de adscripción: División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México.

5.1. Introducción

La diabetes mellitus tipo 2 con complicaciones renales, también conocida como nefropatía diabética, es una complicación que afecta a los riñones de personas con este tipo de diabetes. Lo anterior se debe a que los altos niveles de azúcar en la sangre dañan severamente los vasos sanguíneos de los riñones, afectando la capacidad de desechar las toxinas del cuerpo. La nefropatía diabética es también conocida como enfermedad renal diabética.

El crecimiento exponencial de la diabetes a escala global representa un desafío significativo. Los datos del más reciente Atlas de la Federación Internacional de Diabetes (FID) (International Diabetes Federation, 2025) señalan que el 11.1 % de la población adulta (uno de cada nueve) vive con esta condición. La FID menciona que un factor complicado desde el punto de vista de salud pública es la tasa de infradiagnóstico, dado que más del 40 % de los afectados no han sido identificados La creciente incidencia de la diabetes mellitus tipo 2 en México establece un contexto epidemiológico de alta prioridad para la salud pública. Según los registros de la Dirección General de Epidemiología de la Secretaría de Salud en México (Secretaría de Salud, 2025), hasta el tercer trimestre de 2025 se contabilizó un total de 34 466 ingresos de pacientes con diagnóstico de positivo al sistema. Dentro de este panorama nacional, es notable que el estado de Tabasco se encuentra entre las entidades con el mayor número de casos reportados, junto con Jalisco y Puebla. Esta concentración de pacientes en Tabasco subraya la urgencia de desarrollar herramientas predictivas locales enfocadas en las complicaciones más graves de la enfermedad, como la nefropatía diabética.

Por otro lado, la Inteligencia Artificial (IA), mediante el Aprendizaje automático, permite generar modelos predictivos de forma autónoma. Estos algoritmos aprenden de los datos, reconocen patrones y mejoran sin necesidad de programación explícita, automatizando la creación de modelos estadísticos avanzados y reduciendo la intervención humana (Deitel y Deitel, 2019). En la literatura especializada se puede encontrar que los algoritmos de Aprendizaje automático se pueden clasificar de varias maneras, siendo una de las más novedosas, según su interpretabilidad. En esta clasificación se ubican los algoritmos que producen modelos denominados de “caja blanca” (como los Árboles de decisión o Naïve Bayes) y los algoritmos que producen modelos de “caja negra” (como los Bosques aleatorios), que generalmente ofrecen mejores resultados pero

dificultan comprender sus predicciones (Molnar, 2024).

Este dilema de los métodos de caja negra representa uno de los desafíos centrales en la adopción clínica de la IA. Por ello, para fortalecer la confianza de los usuarios y garantizar la transparencia en las predicciones complejas, se hace esencial el uso de técnicas de interpretabilidad. En este estudio, se utiliza la herramienta LIME (*Local Interpretable Model-agnostic Explanations*) para abordar este problema, ya que permite ofrecer explicaciones locales del comportamiento de cualquier modelo, detallando la contribución de cada variable a una predicción individual (Ribeiro et al., 2016).

5.2. Materiales y métodos

5.2.1. Conjunto de datos

En este estudio se utilizó el dataset denominado DiabetIA, generado a partir de la información registrada en el Sistema de Informática de Medicina Familiar (SIMF) del IMSS, en el estado de Michoacán (Instituto Mexicano del Seguro Social, 2025). La base de datos se conformó mediante la consulta electrónica de expedientes clínicos de pacientes diagnosticados con diabetes mellitus tipo 2, siguiendo los criterios de la Clasificación Internacional de Enfermedades (CIE). El dataset integra 42,182 instancias y 520 variables, dentro de las cuales se prestó especial atención a la variable `e112`, que corresponde a pacientes con diabetes tipo 2 que presentan complicaciones renales y constituye el foco principal de este análisis.

Por otro lado, existe un dataset público llamado *Chronic Kidney Disease* (CKD) que contiene datos de pacientes con insuficiencia renal crónica originario de un hospital de la India. Este conjunto de datos se encuentra alojado en el *UC Irvine Machine Learning Repository*, y consta de 400 instancias y 23 variables, incluyendo además la clase, que indica paciente con insuficiencia renal o paciente sin esta condición (Rubini et al., 2015). Este dataset ha sido trabajado por múltiples autores a nivel internacional, creando modelos predictivos utilizando diversos algoritmos de aprendizaje automático y obteniendo resultados satisfactorios (Hernández-Gaspar et al., 2025). Por esta razón, en este estudio se seleccionaron las 23 variables del dataset DiabetIA con características similares a las empleadas en el dataset CKD, con el objetivo de crear un modelo

predictivo de rendimiento similar. Las variables consideradas son:

1. `age_at_wx`: Edad del paciente al final del periodo (años).
2. `edad`: Edad del paciente (años).
3. `essential_(primary)_hypertension`: Hipertensión esencial (primaria) (0,1).
4. `fn_ta_systolic_mean`: Valor medio de la presión sistólica (mm/Hg).
5. `fn_ta_diastolic_mean`: Valor medio de la presión diastólica (mm/Hg).
6. `cs_sex`: Sexo (F,M).
7. `obesity`: Obesidad (0,1).
8. `fn_ego_mean`: Valor medio del análisis de la orina (Mg/dL).
9. `fn_ph_mean`: Valor medio del potencial de hidrógeno en la sangre (0–14).
10. `fn_density_mean`: Valor medio de la densidad urinaria (g/ml).
11. `hear_failure`: Insuficiencia cardiaca (0,1).
12. `fn_weight_mean`: Valor medio del peso (kg).
13. `fn_height_mean`: Valor medio de la altura (m).
14. `fn_hemoglobin_mean`: Valor medio de la hemoglobina (gms).
15. `fn_glycemia_mean`: Valor medio de la glucosa en la sangre (Mg/dL).
16. `fn_right_foot_mean`: Índice médico del pie diabético derecho (0.5–1.5).
17. `fn_left_foot_mean`: Índice médico del pie diabético izquierdo (0.5–1.5).
18. `other_fluid_electrolyte_and_acid_base_disorders`: Otros trastornos del equilibrio hidroelectrolítico y ácido-base (0,1).
19. `protein_calorie_malnutrition_unspecified`: Desnutrición proteico-calórica no especificada (0,1).
20. `fn_fasting_glucose_mean`: Valor medio de glucosa en ayunas (Mg/dL).
21. `fn_aurico_mean`: Valor medio del ácido úrico en la orina (Mg/dL).
22. `e112`: Variable clase, Diabetes tipo 2 con complicaciones renales (0=Sin complicaciones, 1=Nefropatía diabética).

Una vez seleccionadas las variables, el preprocesamiento de datos constituye una etapa fundamental para garantizar la calidad, estabilidad y confiabilidad de los modelos predictivos a crear. A continuación se describen los pasos aplicados para la depuración y preparación del conjunto de datos.

5.2.2. Análisis inicial y detección de valores faltantes

En primera instancia se evaluó la cantidad de valores faltantes en el dataset; se calculó la proporción de valores faltantes por variable, revelando que varias de ellas presentaban niveles elevados de valores nulos. Con el fin de evitar ruido estadístico, sesgos en la etapa de imputación y un deterioro significativo del rendimiento de los modelos, se decidió eliminar las variables que presentaban más del 40 % de valores faltantes. Este umbral se seleccionó siguiendo recomendaciones de la literatura, donde niveles superiores al 30–50 % comprometen la validez de la variable y afectan negativamente tanto la estabilidad como la interpretabilidad del modelo (Little y Rubin, 2019). Después de aplicar este preprocesamiento, el dataset quedó reducido a 13 variables con niveles aceptables de valores nulos, todas ellas clínicamente relevantes y con patrones de ausencia manejables.

5.2.3. Imputación de valores faltantes

Para conservar la integridad del conjunto de datos y evitar pérdidas innecesarias de información, se aplicaron métodos de imputación dependiendo el tipo de dato de cada variable:

- Variables numéricas: imputadas con la media de cada atributo.
- Variables categóricas: imputadas utilizando la moda (valor más frecuente).

Estos métodos fueron seleccionados debido a la distribución relativamente uniforme de las variables conservadas y porque son compatibles con algoritmos interpretables tales como LIME.

5.2.4. Estandarización y normalización

La codificación de variables categóricas convierte etiquetas en representaciones numéricas, evitando que el modelo interprete categorías nominales como valores ordinales. Este paso asegura que la información cualitativa pueda ser procesada de manera adecuada por cualquier algoritmo de Aprendizaje automático. La estandarización de variables numéricas ajusta los datos a una misma escala, con media cero y desviación estándar uno. Esto previene que atributos con magnitudes mayores dominen el entrenamiento y favorece un aprendizaje más equilibrado y preciso.

5.2.5. Balanceo del conjunto de datos

El análisis inicial mostró un fuerte desbalance de clases: 41 088 pacientes sin nefropatía diabetíca versus 1,092 pacientes con complicaciones renales. Este desbalance podría sesgar los modelos hacia la clase mayoritaria, reduciendo la capacidad para identificar casos positivos. Para eliminar este problema se aplicó la técnica de submuestreo aleatorio (*undersampling*), equilibrando el número de muestras de ambas clases.

5.3. Algoritmos de Aprendizaje automático

En este estudio se emplearon distintos algoritmos de aprendizaje automático, tanto tradicionales como uno interpretable, los cuales se describen a continuación.

5.3.1. Algoritmos clásicos

Los algoritmos considerados tradicionales de IA son esenciales en tareas de predicción y clasificación. Entre ellos, destacan los Árboles de decisión ya que segmentan el espacio de características mediante divisiones recursivas, lo que facilita la interpretación en problemas de clasificación y regresión (Breiman et al., 1984). Las máquinas de vector soporte (SVM) optimizan un hiperplano que maximiza el margen entre clases, abordando problemas lineales y no lineales mediante funciones kernel (Cortes y Vapnik, 1995; Poveda García et al., 2024). El Naïve Bayes, basado en el teorema de Bayes, asume independencia condicional entre predictores, siendo eficiente para grandes volúmenes de datos (Maron, 1961). Por su parte, el Gradient boosting combina secuencialmente modelos débiles para minimizar una función de pérdida mediante gradientes (Friedman, 2001). Finalmente, Random forest construye múltiples árboles sobre subconjuntos aleatorios y agrega sus resultados para mejorar la precisión y reducir el sobreajuste (Salman et al., 2024).

5.3.2. Algoritmo interpretable

En el contexto del aprendizaje automático, la interpretabilidad y la explicabilidad son esenciales para comprender cómo los modelos generan predicciones. La interpretabilidad se refiere

al grado en que un humano puede entender el funcionamiento interno del modelo y la relación entre las variables de entrada y salida, ya sea de forma global (comportamiento general) o local (decisiones específicas). Por su parte, la explicabilidad se centra en ofrecer justificaciones comprensibles sobre las predicciones, lo que puede lograrse mediante técnicas externas que aportan claridad a modelos complejos o de tipo “caja negra”.

Por lo anterior, en este trabajo se empleó un enfoque que permite analizar el comportamiento de los modelos generados. LIME aproxima el funcionamiento de algoritmos complejos mediante representaciones locales simples alrededor de una predicción, lo que facilita comprender cómo cada variable contribuye en casos específicos y aporta transparencia al proceso de decisión (Ribeiro et al., 2016).

Para el análisis y modelado se utilizó el ecosistema de bibliotecas de Python, destacando la biblioteca Scikit-learn, ampliamente empleada en aprendizaje automático por su soporte a algoritmos supervisados, además de utilidades para preprocesamiento, selección de modelos, ajuste de hiperparámetros y evaluación (Pedregosa et al., 2011).

Para la manipulación y limpieza de datos estructurados se utilizó la biblioteca Pandas (McKinney, 2010), mientras que Matplotlib (Hunter, 2007) y Seaborn (Waskom, 2021) facilitaron la generación de visualizaciones estadísticas para explorar patrones y relaciones. La implementación se llevó a cabo en Google Colab, un entorno basado en *notebooks* Jupyter que ofrece recursos computacionales escalables y reproducibles (Carneiro et al., 2018).

5.4. Métricas de evaluación

La matriz de confusión (Erbani et al., 2024) es una herramienta que resume los conteos de predicciones correctas e incorrectas. En ella se representan los verdaderos positivos (TP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (TN). Esta matriz permite visualizar el rendimiento del modelo y sirve como base para calcular métricas derivadas como exactitud, precisión y sensibilidad.

La exactitud (*Accuracy*) se refiere a la proporción de predicciones correctas, tanto positivas como negativas, respecto al total de observaciones (Mengle y Gurmendez, 2019). La cual se calcula en la Ec. (6.7) y es una medida global del desempeño del modelo, aunque puede resultar

poco informativa cuando los datos están desbalanceados.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5.1}$$

La precisión (*Precision*) mide la proporción de instancias clasificadas como positivas que realmente lo son (Cohen et al., 2024). La cual se define en la Ec. (6.8), y es útil para minimizar falsos positivos; resulta importante en contextos donde una clasificación errónea positiva tiene un costo elevado.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

La sensibilidad (*Recall*) (Cohen et al., 2024) indica la proporción de positivos reales que fueron correctamente identificados por el modelo. Se calcula con la Ec. (6.9) y es crucial en situaciones donde no detectar un positivo es más costoso, como en diagnósticos médicos.

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

5.5. Trabajos relacionados

Existe investigación relacionada con la creación de modelos interpretables empleando LIME, particularmente utilizando el dataset CKD mencionado anteriormente (Gaur et al., 2022b; S. K. Ghosh y Khandoker, 2024). Otros trabajos han empleado LIME para mejorar la interpretabilidad de modelos en entornos médicos. En la Tabla 5.1 se muestra un concentrado de trabajos, los cuales destacan la utilidad de los modelos interpretables para aumentar la confianza en aplicaciones críticas.

5.6. Proceso experimental

Los datos se dividieron en conjuntos de entrenamiento y prueba en una proporción de $\frac{2}{3}$ y $\frac{1}{3}$, respectivamente. Posteriormente, se entrenaron diversos modelos de aprendizaje automático, entre ellos Árboles de decisión, SVM, Naïve Bayes, Gradient boosting y Random forest. La

Tabla 5.1. Estudios recientes que emplean LIME en modelos orientados al área médica.

Trabajo	Técnica	Mejor modelo	Rendimiento
Kalusivalingam et al., 2021	LIME, SHAP	XGBoost	93.1 %
Gaur et al., 2022b	LIME, SHAP, DALEX	EBM	98.57 %
Y. Wu et al., 2023	LIME	Random Forest	94.2 %
Shaikh et al., 2023	LIME, SHAP	SVM	90.5 %
S. K. Ghosh y Khandoker, 2024	LIME, SHAP	XGBoost	93.29 %
Paul, 2024	LIME, SHAP	Gradient Boosting	92.8 %
Rahimiaghdam y Alemdar, 2025	MindfulLIME	–	Sin métrica (explicaciones estables)

etapa final correspondió a la interpretación de los resultados mediante LIME. Este procedimiento garantiza un adecuado manejo de los datos y la solidez de los modelos, permitiendo obtener predicciones confiables y transparentes.

5.7. Resultados y discusión

5.7.1. Algoritmos de Aprendizaje automático

En la Tabla 5.2 se presentan los resultados obtenidos, evaluados mediante las métricas de exactitud, precisión y sensibilidad, con el objetivo de comparar tanto su rendimiento como su interpretabilidad.

Tabla 5.2. Matrices de confusión por algoritmo. Mejores resultados en **negritas**.

Clasificador	TP	FP	FN	TN
Naïve Bayes Gaussiano	186	32	70	149
SVM	191	**27**	79	140
Árboles de decisión	151	67	**55**	**164**
Gradient boosting	187	31	66	153
Random forest	**195**	28	69	145

El modelo Random forest obtuvo el mejor desempeño global con una precisión del 80 %, seguido por Gradient boosting (78 %). En términos de sensibilidad, el SVM alcanzó el valor más alto (0.88), lo que indica una mayor capacidad para identificar correctamente las instancias positivas. Por otro lado, Árboles de decisión mostró el rendimiento más bajo en exactitud y sensibilidad, lo

que sugiere menor estabilidad frente a los demás modelos.

En la Tabla 5.3 el modelo Random forest muestra la mayor cantidad de verdaderos positivos (TP = 195) y verdaderos negativos (TN = 145), lo que confirma su superioridad en precisión global. Por otro lado, SVM presenta el menor número de falsos positivos (FP = 27), lo que indica una alta especificidad. En contraste, Árboles de decisión exhibe la mayor cantidad de errores (FP = 67), reflejando menor estabilidad. En general, los modelos basados en ensambles (Random forest y Gradient boosting) ofrecen un mejor equilibrio entre sensibilidad y precisión.

Tabla 5.3. Matrices de confusión por algoritmo. Mejores resultados en **negritas**.

Clasificador	**TP**	**FP**	**FN**	**TN**
Naïve Bayes	186	32	70	149
SVM	191	**27**	79	140
Árboles de decisión	151	67	**55**	**164**
Gradient boosting	187	31	66	153
Random forest	**195**	28	69	145

5.7.2. Interpetabilidad con LIME

LIME es una técnica de interpretabilidad local orientada a examinar la contribución individual de cada característica en las predicciones del modelo. El propósito es identificar las variables con mayor impacto en la clasificación de cada paciente, ofreciendo una explicación comprensible sobre las decisiones del modelo.

Debido a que Random forest obtuvo el mejor rendimiento, se tomó como base para LIME y así poder interpretar los resultados de forma local, mostrando cómo cada atributo influye positiva o negativamente en la predicción final. En ese sentido, se analizan casos representativos en los que el modelo acierta (TP, verdaderos positivos y TN, verdaderos negativos) y cuando falla (FP, falsos positivos y FN, falsos negativos).

Los valores entre paréntesis (positivo o negativo) representan la influencia que cada característica tuvo en la predicción del modelo según el análisis local de LIME.

- Un signo positivo indica que la característica favoreció la clase predicha por el modelo.
- Un signo negativo señala que la característica presionó la predicción hacia la clase contraria.

Estos valores se obtienen mediante una regresión lineal ajustada localmente por LIME en torno al caso analizado, y reflejan el peso individual de cada variable en la decisión del modelo.

Caso 1: Predicción correcta (TN). El modelo Random forest predijo correctamente que el paciente no presenta complicaciones renales (clase 0), coincidiendo con la etiqueta real. La probabilidad asignada a la clase 0 fue de 1.00, lo que indica una confianza total en la predicción. El análisis con LIME muestra que las características con mayor influencia positiva hacia la esta clase son:

- `essential_(primary)_hypertension` = 0 (+0.23): la ausencia de hipertensión primaria contribuye fuertemente a que la predicción sea *Sin complicación*.
- `age_at_wx` (+0.20): la edad al momento del diagnóstico influye en el diagnóstico *Sin complicación*.
- `edad` (+0.09): la edad actual del paciente influye también en el mismo diagnóstico.
- `fn_systolic_mean` (+0.07): la presión sistólica influye ligeramente hacia el diagnóstico.
- `cs_sex_M` (+0.04): el sexo del paciente influye ligeramente también en el mismo diagnóstico.

Este resultado evidencia que la ausencia de hipertensión y la edad relativamente baja son factores determinantes para la clasificación correcta (Figura 5.1).

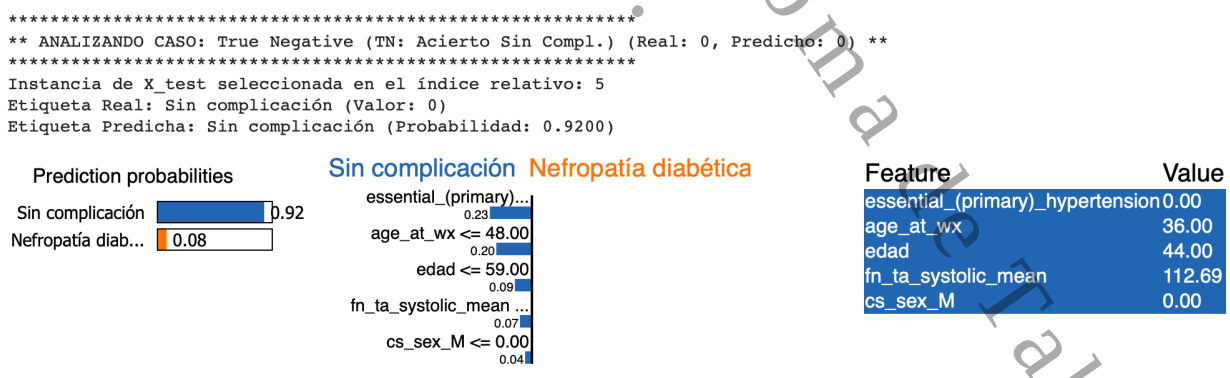

Figura 5.1. Resultado de LIME: paciente con diagnóstico correcto (TN).

Caso 2: Predicción correcta (TP). En este caso, el modelo predijo correctamente que el paciente presenta complicaciones renales (Etiqueta real: 1, Predicción: 1). La probabilidad asignada por el modelo fue de 0.66, lo que indica una confianza moderada en la clasificación. Las características con mayor influencia según LIME fueron:

- `essential_primary_hypertension` (+0.23): la presencia de hipertensión primaria fue el factor más determinante para la clase con *Nefropatía diabética*.
- `age_at_wx` (+0.12): la edad al momento del examen mayor a 70 años incrementó la probabilidad de *Nefropatía diabética*.
- `edad` (0.03): la edad actual superior a 80 años reforzó la predicción.
- `fn_ta_systolic_mean` (+0.06): el promedio de presión sistólica tuvo influencia hacia la clase *Nefropatía diabética*.
- `cs_sex_M` (-0.03): el sexo masculino aportó un efecto mínimo hacia la clase *Sin complicación*.

El modelo considera la hipertensión y la edad avanzada como factores clave para clasificar al paciente con complicaciones renales. Aunque la probabilidad no es alta (0.66), la explicación es coherente con criterios clínicos, lo que sugiere que el modelo identifica correctamente patrones de riesgo (ver Figura 5.2).

```
**********************************************************
** ANALIZANDO CASO: True Positive (TP: Acierto Con Compl.) (Real: 1, Predicho: 1) **
**********************************************************
Instancia de X_test seleccionada en el índice relativo: 2
Etiqueta Real: Nefropatía diabética (Valor: 1)
Etiqueta Predicha: Nefropatía diabética (Probabilidad: 0.9600)
```

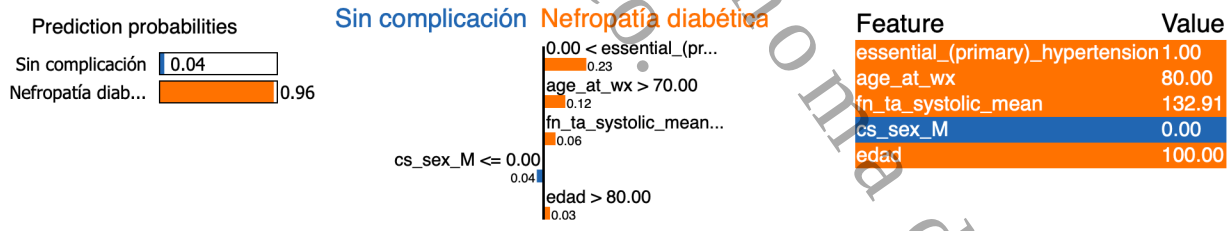

Figura 5.2. Resultado de LIME: paciente con diagnóstico correcto (TP).

Caso 3: Predicción incorrecta (FP). En este caso, el modelo predijo que el paciente presentaba complicaciones renales (Predicción: 1) cuando en realidad no las tenía (Etiqueta real: 0). La probabilidad asignada fue de 0.80, lo que indica una alta confianza en esta predicción errónea. Las características con mayor influencia fueron:

- `essential_primary_hypertension` (+0.24): la presencia de hipertensión primaria fue el factor más determinante para la clase *Nefropatía diabética*.

- `age_at_wx` (+0.11): la edad al momento del examen mayor a 70 años incrementó la probabilidad de *Nefropatía diabética*.
- `fn_weight_mean_mean` (+0.02): el promedio del peso aportó influencia positiva hacia la clase con *Nefropatía diabética*.
- `edad` (+0.04): la edad actual superior a 80 años reforzó la predicción.
- `cs_sex_M` (-0.04): el sexo masculino tuvo un impacto mínimo hacia la clase con *Nefropatía diabética*.

El modelo sobreestimó el riesgo debido a factores clínicos como hipertensión y edad avanzada, lo que llevó a un falso positivo. Aunque clínicamente no es tan grave como un falso negativo, este tipo de error puede generar intervenciones innecesarias y costos adicionales (Figura 5.3).

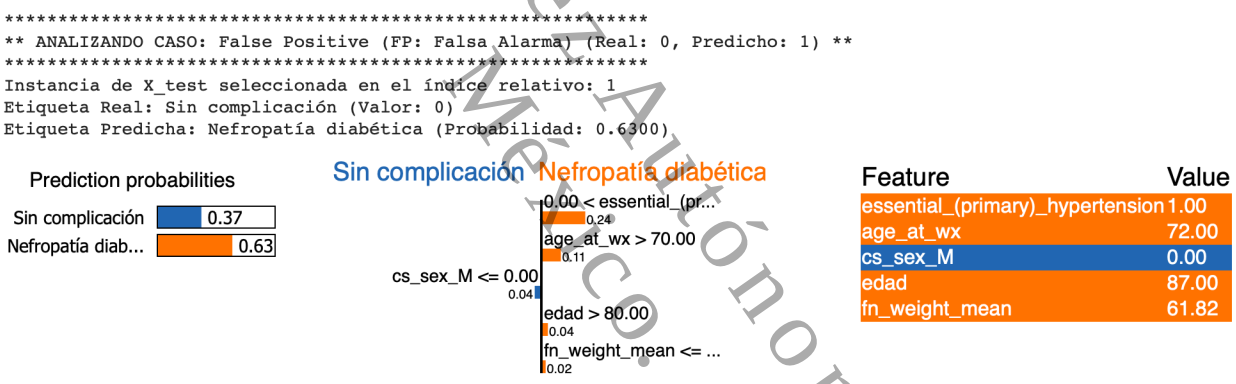

Figura 5.3. Resultado de LIME: paciente con diagnóstico incorrecto (FP).

Caso 4: Predicción incorrecta (FN). Este caso representa el error más grave desde el punto de vista clínico. La predicción del modelo indica que el paciente no presenta complicaciones renales (Predicción: 0) cuando en realidad sí las presenta (Etiqueta real: 1). La probabilidad asignada fue de 0.66 para la clase *Sin Complicación*, lo que indica una confianza moderada en esta predicción incorrecta. Las características con mayor influencia fueron:

- `essential_primary_hypertension` (+0.24): la ausencia de hipertensión primaria contribuyó fuertemente hacia la clase *Sin complicación*.
- `edad` (-0.02): la edad aportó poca influencia hacia la clase *Nefropatía diabética*.

- `fn_ta_systolic_mean` (+0.02): el promedio de presión sistólica tuvo un efecto positivo hacia la clase *Sin complicación*.
- `age_ta_wx` (-0.04): el promedio de la edad al inicio también tuvo un efecto hacia la clase con *Nefropatía diabética*.
- `cs_sex_M` (+0.04): el sexo también tuvo un leve efecto hacia la clase *Sin complicación*.

El modelo subestimó el riesgo debido a la ausencia de hipertensión y valores de presión arterial que no fueron considerados críticos, lo que llevó a un falso negativo. Este tipo de error es clínicamente peligroso, ya que implica no detectar una condición que requiere atención, aumentando el riesgo de complicaciones graves (ver Figura 5.4).

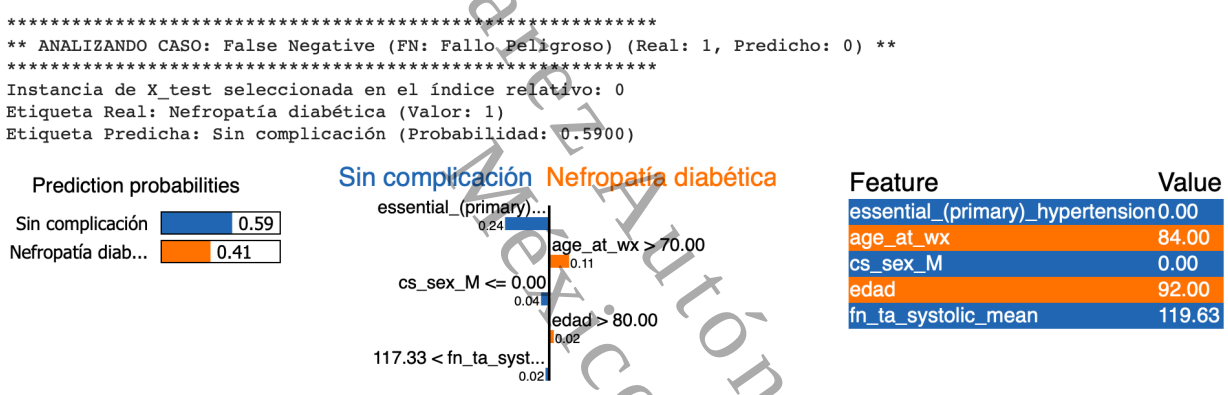

Figura 5.4. Resultado de LIME: paciente con diagnóstico incorrecto (FN).

5.8. Conclusiones

En este trabajo se hace uso de algoritmos de Aprendizaje automático para la clasificación de pacientes con nefropatía diabética, de los cuales el algoritmo de Random forest obtuvo el mejor rendimiento de entre todos los algoritmos seleccionados.

El análisis realizado con el método LIME permitió comprender cómo las características individuales influyen en las predicciones del modelo, ofreciendo transparencia en algoritmos robustos aunque de caja negra, como Random forest. Los casos evaluados muestran que el modelo tiende a basar sus decisiones en factores clínicos relevantes, como hipertensión y edad, lo que coincide con criterios médicos esperados.

Sin embargo, se identificaron errores críticos que evidencian riesgos clínicos importantes. Los falsos positivos, aunque generan intervenciones innecesarias, son menos graves que los falsos negativos, donde el modelo no detecta una condición real, lo que puede retrasar el tratamiento y aumentar el riesgo para el paciente. Este hallazgo subraya la necesidad de complementar los modelos predictivos con mecanismos de validación clínica y estrategias para reducir errores de clasificación.

En este caso de estudio, LIME demostró ser una herramienta útil para evaluar la coherencia del modelo y detectar posibles sesgos, contribuyendo a mejorar la confianza y seguridad en sistemas de apoyo a la decisión médica.

Referencias del capítulo

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth; Brooks/Cole.

Carneiro, T., Medeiros, R. M., Nepomuceno, D. B., Bian, G. B., da Costa, V. H., & de Albuquerque, P. S. (2018). Performance Analysis of Google Colaboratory as a Tool for Machine Learning. *Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1234-1239. https://doi.org/10.1109/CSCI46756.2018.00235

Cohen, L., Mansour, Y., Moran, S., & Shao, H. (2024). Probably Approximately Precision and Recall Learning [Disponible en: https://arxiv.org/abs/2411.13029]. *arXiv preprint arXiv:2411.13029*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Deitel, P., & Deitel, H. (2019). *Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and the Cloud* (1.ª ed.). Pearson.

Erbani, J., Portier, P.-É., Egyed-Zsigmond, E., & Nurbakova, D. (2024). Confusion Matrices: A Unified Theory. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2024.10769075

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gaur, L., Biswas, M., Bakshi, S., Gupta, P., Si, T., Mallik, S., & Maulik, U. (2022b). An Integrated Model to Evaluate the Transparency in Predicting Chronic Kidney Disease Using a Trio-Embedded Explainable Model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4129888

Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, *14*(3687). https://doi.org/10.1038/s41598-024-54375-4

Hernández-Gaspar, S. J., Hernández-Ocaña, B., Hernández-Torruco, J., & Chávez-Bosquez, O. (2025). Inteligencia artificial explicable para el análisis de pacientes con insuficiencia renal crónica. *Nova Scientia*, *17*, 1-22. https://doi.org/10.21640/ns.v17iSpecialIa.3630

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Instituto Mexicano del Seguro Social. (2025). DiabetIA: Conjunto de datos generado a partir del Sistema de Informática de Medicina Familiar (SIMF) en Michoacán [Información derivada de expedientes clínicos de pacientes con diabetes mellitus tipo 2, siguiendo criterios de la Clasificación Internacional de Enfermedades (CIE)].

International Diabetes Federation. (2025). *IDF Diabetes Atlas — Datos y cifras sobre la diabetes* [Accedido: 19-nov-2025]. International Diabetes Federation. https://idf.org/es/about-diabetes/diabetes-facts-figures/

Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2021). Leveraging SHAP and LIME for Enhanced Explainability in AI-Driven Diagnostic Systems. *International Journal of AI and ML*, *2*(3). https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/81

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3.ª ed., Vol. 793). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Maron, M. E. (1961). Automatic indexing: An experiment with a statistical model. *Journal of the ACM*, *8*(3), 404-417. https://doi.org/10.1145/321075.321084

McKinney, W. (2010). Data Structures for Statistical Computing in Python. En S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a

Mengle, S. S., & Gurmendez, M. (2019). *Mastering Machine Learning on AWS*. Packt Publishing.

Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book

Paul, J. (2024). LIME and SHAP interpretability for medical AI systems. https://doi.org/10.13140/RG.2.2.14652.45443

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Poveda García, D. A., Ramírez Guzmán, M. E., Fernández Ordóñez, Y. M., & García. López, E. F. (2024). Implementación del método de optimización de máquinas de soporte vectorial en R. *Tecnura*, *28*(80), 99-118. https://doi.org/10.14483/22487638.20326

Rahimiaghdam, S., & Alemdar, H. (2025). MindfulLIME: A Stable Solution for Explanations of Machine Learning Models with Enhanced Localization Precision. *arXiv preprint arXiv:2503.20758*. https://arxiv.org/abs/2503.20758

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You¿': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Dataset. https://doi.org/10.24432/C5G020

Salman, M., Al-Saedi, A., & Al-Saedi, H. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2*(1), 1-10. https://doi.org/10.5281/zenodo.10812345

Secretaría de Salud. (2025). *Informe Trimestral de Vigilancia Epidemiológica Hospitalaria de Diabetes Mellitus Tipo 2 (3er Trimestre 2025)* (Informe Trimestral) (Accedido: 19-nov-2025). Gobierno de México. https://www.gob.mx/cms/uploads/attachment/file/1029886/InformeTrimestralSVEDMT2-3erTrim-2025.pdf

Shaikh, A. S., Samant, R. M., Patil, K. S., Patil, N. R., & Mirkale, A. R. (2023). Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain. *International Journal of Computer Applications*, *185*(45), 18-23. https://www.ijcaonline.org/archives/volume185/number45/32992-2023923263/

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wu, Y., Zhang, L., Bhatti, U. A., & Huang, M. (2023). Interpretable Machine Learning for Personalized Medical Recommendations: A LIME-Based Approach. *Diagnostics*, *13*(16), 2681. https://doi.org/10.3390/diagnostics13162681

Capítulo 6

EBM y SHAP para predicción de nefropatía diabética en pacientes con diabetes tipo 2

Predicción de Nefropatía Diabética en Pacientes con Diabetes Tipo 2 mediante Inteligencia Artificial Explicable

Simón Javier Hernández-Gaspar; Betania Hernández-Ocaña; José Hernández-Torruco; Oscar Chávez-Bosquez

Resumen, Este trabajo aborda la predicción de complicaciones renales en pacientes con Diabetes Mellitus tipo 2 en población mexicana, utilizando el conjunto de datos *DiabetIA*. Ante la necesidad de transparencia en la toma de decisiones médicas, se evaluó la eficacia de la Inteligencia Artificial Explicable (XAI), comparando modelos intrínsecamente interpretables (*Explainable Boosting Machine*, EBM) frente a explicaciones *post-hoc* (SHAP) sobre modelos de "caja negra". Los resultados demuestran que EBM alcanza un rendimiento predictivo competitivo (precisión $\approx$ 80 %), similar a algoritmos complejos como Random forest, pero con la ventaja de ofrecer trazabilidad directa. El análisis dual confirmó que la hipertensión arterial y la edad son los predictores dominantes. Adicionalmente, la auditoría de errores mediante SHAP reveló una vulnerabilidad en los modelos: una tendencia a subestimar el riesgo en pacientes normotensos (falsos negativos), donde la ausencia de hipertensión actúa como un factor de sesgo que oculta otras señales de alarma. Se concluye que la integración de EBM y SHAP no solo mejora la confianza del personal médico, sino que actúa como una herramienta de auditoría indispensable para refinar diagnósticos y reducir omisiones graves en la medicina de precisión.

Palabras clave: Inteligencia Artificial Explicable (XAI), EBM, SHAP, Nefropatía diabética, Medicina de precisión.

Institución de adscripción: División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México.

6.1. Introducción

La Enfermedad Renal Diabética (ERD), tradicionalmente denominada nefropatía diabética, representa una de las complicaciones microvasculares más críticas de la Diabetes Mellitus Tipo 2 (DMT2). Fisiopatológicamente, la exposición prolongada a niveles elevados de glucemia induce daño en la estructura vascular y en las unidades de filtración de los riñones (glomérulos), comprometiendo severamente la homeostasis metabólica y la capacidad de excreción de toxinas (KDIGO, 2022). A nivel global, la prevalencia de esta condición mantiene una tendencia ascendente. De acuerdo con el Atlas de la Federación Internacional de Diabetes (IDF, 2021), aproximadamente el 11 % de la población adulta mundial vive con diabetes, de los cuales un alto porcentaje desconoce su diagnóstico. Esta problemática se agudiza al considerar que, en regiones como América del Norte, uno de cada tres pacientes con diabetes desarrolla complicaciones renales crónicas. En el contexto mexicano, la Encuesta Nacional de Salud y Nutrición (Instituto Nacional de Salud Pública, 2021) estima que 12.4 millones de adultos padecen diabetes, siendo la población mayor de 60 años la más vulnerable.

La relevancia de este estudio se destaca a partir de las estadísticas más recientes reportadas por la Dirección General de Epidemiología (DGE) de México. De acuerdo con el informe del tercer trimestre de 2025 del Sistema de Vigilancia Epidemiológica Hospitalaria de Diabetes Mellitus Tipo 2 (SVEHDMT2), se registraron un total de 34 466 casos de pacientes con complicaciones relacionadas con la diabetes, incluyendo afectación renal, dentro del sistema nacional de vigilancia (Dirección General de Epidemiología, 2025). En este escenario, el estado de Tabasco destaca como una de las regiones más afectadas, situándose entre las entidades con mayor incidencia de casos reportados, junto con Jalisco y Puebla. Esta alta prevalencia local no solo confirma una carga significativa de salud pública en la región, sino que también pone de manifiesto la urgencia de implementar herramientas predictivas basadas en datos clínicos locales que permitan identificar tempranamente la progresión hacia la nefropatía diabética.

Ante el incremento de la prevalencia de nefropatía diabética en el estado de Tabasco, surge la necesidad de implementar herramientas tecnológicas que faciliten la detección temprana y la estratificación de riesgo. En este contexto, el Aprendizaje automático ofrece un paradigma robusto para la construcción de modelos estadísticos que identifican patrones complejos en grandes

volúmenes de datos sin ser programados de forma explícita (Jordan y Mitchell, 2015).

Sin embargo, la adopción de estas tecnologías en el sector salud enfrenta una barrera crítica: la opacidad. Mientras que los modelos de “caja negra” (como los Bosques aleatorios o las Redes neuronales profundas) suelen ofrecer un alto rendimiento predictivo, su funcionamiento interno resulta inescrutable para el clínico. En contraste, los modelos de “caja blanca” o intrínsecamente interpretables garantizan la transparencia necesaria para la toma de decisiones médicas (Molnar, 2024).

Para abordar este desafío, el presente estudio utiliza la librería InterpretML, centrando el análisis en el Explainable Boosting Machine (EBM). El EBM combina la potencia de los modelos basados en *gradient boosting* con la inteligibilidad de los modelos lineales (Nori et al., 2019). Para robustecer la validación, se complementa con la técnica de SHAP (SHapley Additive exPlanations), basada en la teoría de juegos cooperativos, permitiendo descomponer la contribución individual de cada variable clínica tanto a nivel global como en predicciones locales individuales (Lundberg y Lee, 2017a).

6.2. Materiales y métodos

6.2.1. Conjunto de datos

Para este análisis se trabajó con la base de datos DiabetIA, integrada por 520 variables y más de 42 000 instancias de pacientes con diabetes tipo 2 diagnosticados según la Clasificación Internacional de Enfermedades (CIE). Los datos provienen de los expedientes electrónicos del Sistema de Información de Medicina Familiar (SIMF) del IMSS Michoacán. Se puso especial énfasis en la variable `e112` (pacientes con complicaciones renales), debido a su relevancia como objetivo principal de la investigación (Gómez García et al., 2023).

Complementariamente, se tomó como referencia el dataset público Chronic Kidney Disease (CKD) del repositorio UC Irvine (Rubini et al., 2015). Este conjunto de datos, que documenta 23 variables clínicas de pacientes en la India, es un referente internacional donde diversos autores han logrado modelos predictivos con resultados sobresalientes (Hernández-Gaspar et al., 2025). Basándose en el éxito alcanzado en análisis preliminares de esta investigación, se optó por ho-

mologar dicha estructura en el contexto de la población mexicana. En consecuencia, se seleccionaron en DiabetIA 23 atributos con características análogas a los del dataset CKD, conformando así un conjunto de 23 variables (incluyendo la etiqueta de clase) con el objetivo de desarrollar un modelo de alto rendimiento. Las variables consideradas se presentan a continuación:

1. `age_at_wx`: Edad del paciente al final del periodo (años).
2. `edad`: Edad del paciente (años).
3. `essential_(primary)_hypertension`: Hipertensión esencial (primaria) (0,1).
4. `fn_ta_systolic_mean`: Valor medio de la presión sistólica (mm/Hg).
5. `fn_ta_diastolic_mean`: Valor medio de la presión diastólica (mm/Hg).
6. `cs_sex`: Sexo (F,M).
7. `obesity`: Obesidad (0,1).
8. `fn_ego_mean`: Valor medio del análisis de la orina (Mg/dL).
9. `fn_ph_mean`: Valor medio del potencial de hidrógeno en la sangre (0–14).
10. `fn_density_mean`: Valor medio de la densidad urinaria (g/ml).
11. `hear_failure`: Insuficiencia cardiaca (0,1).
12. `fn_weight_mean`: Valor medio del peso (kg).
13. `fn_height_mean`: Valor medio de la altura (m).
14. `fn_hemoglobin_mean`: Valor medio de la hemoglobina (gms).
15. `fn_glycemia_mean`: Valor medio de la glucosa en la sangre (Mg/dL).
16. `fn_right_foot_mean`: Índice médico del pie diabético derecho (0.5–1.5).
17. `fn_left_foot_mean`: Índice médico del pie diabético izquierdo (0.5–1.5).
18. `other_fluid_electrolyte_and_acid_base_disorders`: Otros trastornos del equilibrio hidroelectrolítico y ácido-base (0,1).
19. `protein_calorie_malnutrition_unspecified`: Desnutrición proteico-calórica no especificada (0,1).
20. `fn_fasting_glucose_mean`: Valor medio de glucosa en ayunas (Mg/dL).
21. `fn_creatinine_mean`: Valor medio de la creatina (Mg/dL).
22. `fn_aurico_mean`: Valor medio del ácido úrico en la orina (Mg/dL).
23. `e112`: Variable clase, Diabetes tipo 2 con complicaciones renales (0=Sin complicaciones, 1=Nefropatía diabética).

6.2.2. Preprocesamiento de datos

En la fase de preprocesamiento, se detectó una cantidad significativa de valores ausentes en diversas variables del conjunto de datos como se puede ver en la Figura (6.1). Para preservar la integridad de la información y minimizar el ruido que pudiera comprometer la estabilidad de los modelos, se descartaron las variables con una tasa de valores faltantes superior al 40 % (Fernández et al., 2018). Tras esta depuración, se conservaron 13 variables que cumplen con los criterios de completitud necesarios para garantizar la robustez del análisis predictivo ver Figura (6.2).

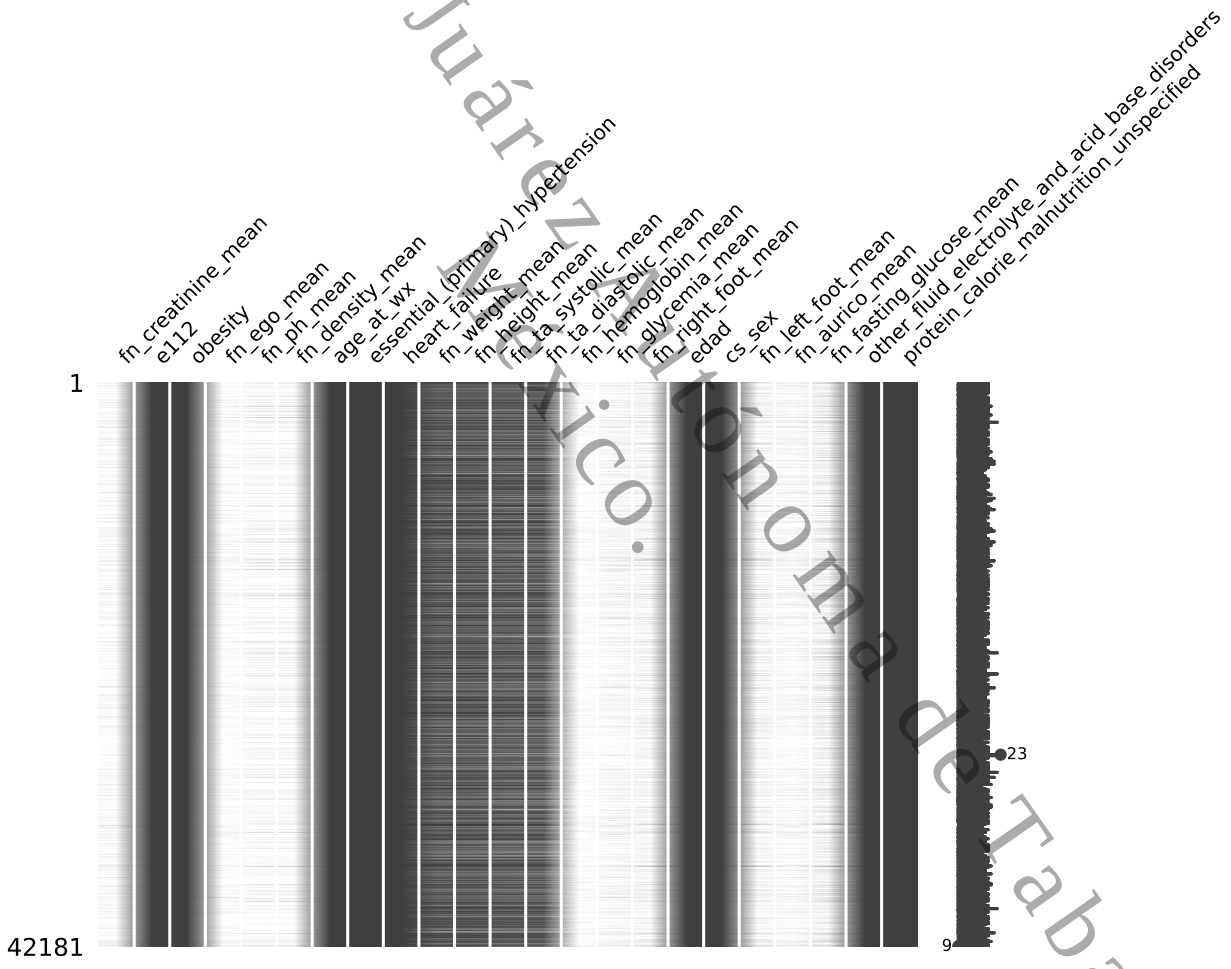

Figura 6.1. Variables sin depurar.

Con el fin de garantizar la calidad y coherencia del conjunto de datos antes del entrenamiento de los modelos, se llevó a cabo una etapa de preprocesamiento estructurada en tres fases, que

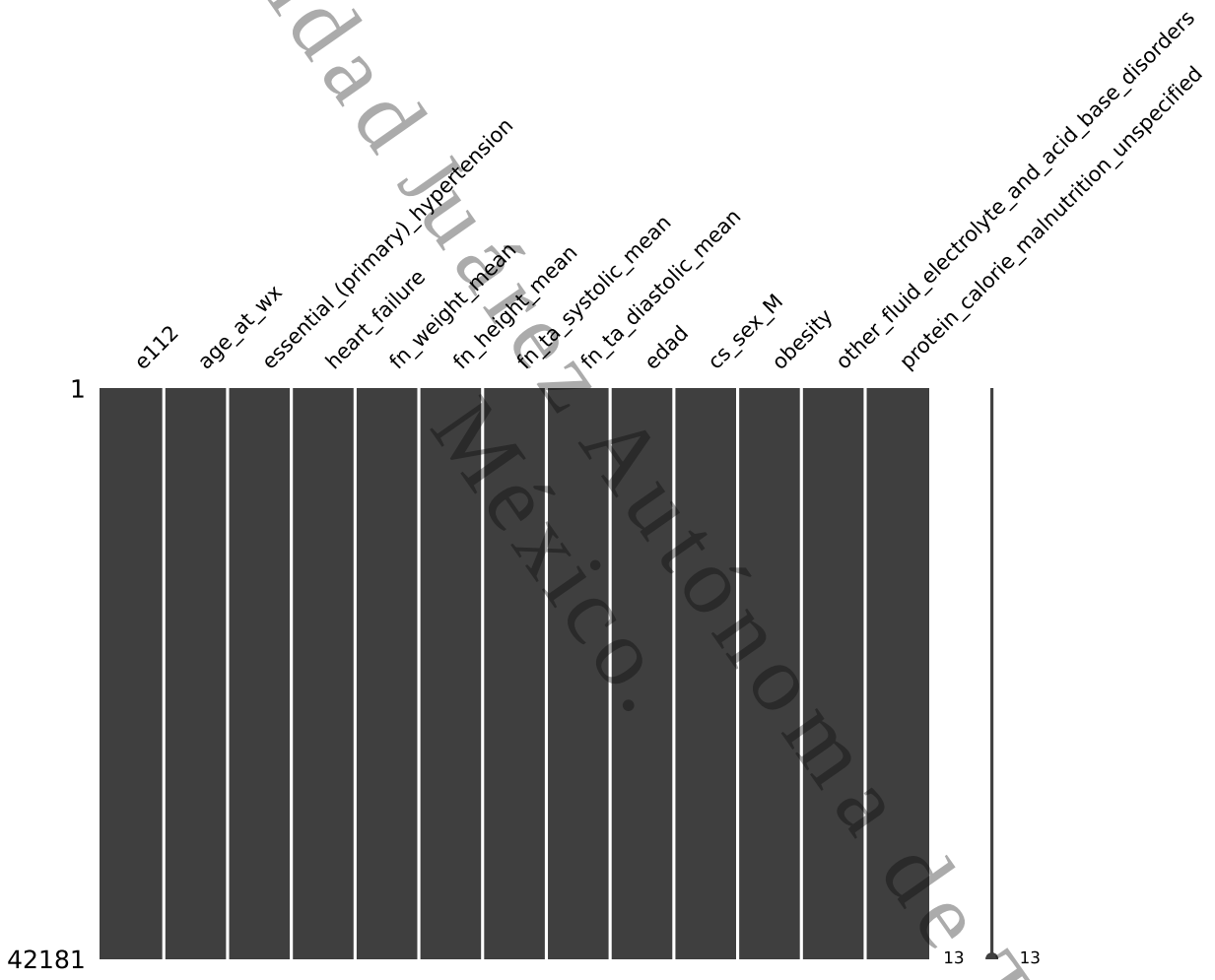

Figura 6.2. Variables filtradas.

se mencionan a continuación.

6.2.3. Imputación de valores faltantes

Los valores ausentes en las variables cuantitativas fueron reemplazados mediante la media aritmética, mientras que para los atributos categóricos se empleó la moda, asegurando así la integridad de la muestra sin introducir sesgos significativos (Kuhn y Johnson, 2013).

6.2.4. Codificación de variables

Se aplicaron técnicas de codificación (*encoding*) a los atributos nominales, transformándolos en representaciones numéricas compatibles con los requisitos de los algoritmos de Aprendizaje automático.

6.2.5. Normalización y escalamiento

Finalmente, se implementaron procesos de transformación y escalado de datos. Esta homogenización de magnitudes previene que las variables con rangos mayores dominen injustificadamente sobre las demás (Géron, 2019) optimizando así la convergencia y el desempeño de los modelos predictivos.

6.2.6. Balanceo del conjunto de datos

Finalmente, se observó que el conjunto de datos presentaba un marcado desbalance de clases, con 41,088 registros correspondientes a otros padecimientos y 1 092 a pacientes con diabetes mellitus tipo 2 con complicaciones renales. Para corregir este desequilibrio y mejorar la capacidad del modelo para detectar casos positivos, se implementó la técnica de submuestreo (*undersampling*) sobre la clase mayoritaria (He y Garcia, 2009) logrando un conjunto de datos balanceado y adecuado para el entrenamiento de los modelos, resultando 2184 instancias y 13 variables.

6.3. Algoritmos de Inteligencia Artificial

Esta investigación emplea un enfoque comparativo que abarca desde algoritmos de clasificación tradicionales hasta técnicas de Inteligencia Artificial Explicable (XAI). Esta selección busca equilibrar la precisión predictiva con la transparencia diagnóstica, permitiendo no solo identificar pacientes con riesgo renal, sino también proporcionar información clínica accionable sobre los factores de riesgo. A continuación, se detallan los algoritmos utilizados.

6.3.1. Algoritmos clásicos

Tabla 6.1. Comparación de Algoritmos de Aprendizaje automático.

Algoritmo	Fundamento	Ventaja clínica	Referencia
Árboles de Decisión	Partición recursiva del espacio de características.	Estructura lógica similar al diagnóstico médico.	Breiman et al., 1984
SVM	Maximización del margen mediante hiperplanos.	Eficacia en fronteras de decisión complejas.	Cortes y Vapnik, 1995
Naïve Bayes	Probabilidad condicional con supuesto de independencia.	Eficiencia computacional en grandes volúmenes.	Maron, 1961
Gradient Boosting	) Optimización secuencial de funciones de pérdida.	Alta capacidad predictiva mediante ensamble.	Friedman, 2001
Random Forest	Ensamble de árboles con *bagging*.	Reducción de varianza y robustez ante *overfitting*.	Breiman, 2001

Previo al modelado, se realizó un Análisis Exploratorio de Datos (EDA) para identificar distribuciones, correlaciones y valores atípicos que pudieran sesgar las predicciones. El rendimiento de los modelos se evaluó mediante una matriz de confusión para sintetizar los conteos de predicciones correctas e incorrectas (Erbani et al., 2024) permitiendo el cálculo de métricas robustas para el contexto clínico. Se calculó la Exactitud (*Accuracy*) como medida global de desempeño (Mishra, 2021), complementada con la Precisión y la Sensibilidad (*Recall*) (Cohen et al., 2024). Estas últimas resultan críticas para balancear el costo de los falsos positivos frente a la necesidad de detectar oportunamente a pacientes con riesgo de insuficiencia renal.

Para el desarrollo de este proyecto se empleó el ecosistema de bibliotecas de Python, destacando Scikit-learn, una biblioteca de código abierto ampliamente utilizada en aprendizaje automático. Esta herramienta proporciona utilidades para el preprocesamiento de datos, la selección

de modelos, el ajuste de hiperparámetros y la evaluación del rendimiento (Pedregosa et al., 2011). Complementariamente, se utilizó la biblioteca Pandas para la manipulación, limpieza y análisis de datos estructurados, mientras que las bibliotecas Matplotlib y Seaborn fueron aplicadas para la construcción de gráficos y visualizaciones estadísticas que permitieron explorar el comportamiento de las variables y detectar patrones relevantes. El análisis se llevó a cabo en la plataforma Google Colab, la cual provee un entorno computacional basado en máquinas virtuales con soporte para Notebooks Jupyter, facilitando la implementación de los métodos y la gestión de los experimentos en un ambiente reproducible y escalable (Bisong, 2019).

6.3.2. Algoritmo interpretable

La adopción de Inteligencia Artificial en salud exige un equilibrio entre precisión y transparencia diagnóstica. En este estudio, la interpretabilidad se aborda como la capacidad de comprender la relación lógica entre variables (Rudin, 2019), mientras que la explicabilidad proporciona justificaciones sobre las decisiones del modelo (Barredo Arrieta et al., 2020).

Para el análisis, se combinaron dos enfoques: Explainable Boosting Machines (EBM) y SHAP. EBM, basado en modelos aditivos generalizados (GAM), ofrecen una transparencia intrínseca capaz de capturar relaciones no lineales sin degradar el rendimiento predictivo (Nori et al., 2019). Por otro lado, SHAP (SHapley Additive exPlanations) aplica la teoría de juegos cooperativos para asignar valores de importancia equitativos a cada característica clínica (Lundberg y Lee, 2017a). Esta integración permite identificar con claridad los factores de riesgo de insuficiencia renal, facilitando la toma de decisiones informada en el ámbito clínico.

6.3.2.1. Modelos Aditivos Explicables (EBM)

Explainable Boosting Machine (EBM) es un modelo de caja de cristal cuya principal ventaja es que la contribución de cada variable a la predicción final es exacta y calculada de forma independiente. La formulación matemática de un modelo EBM se define como:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j) \tag{6.1}$$

Donde g es la función de enlace (*logit* para clasificación binaria), β_0 representa el intercepto, $f_j(x_j)$ es la función de forma (*shape function*) del efecto principal, y $f_{ij}(x_i, x_j)$ representa las interacciones por pares.

En el contexto de la predicción de riesgo, el modelo opera internamente en el espacio de *log-odds*. La naturaleza aditiva justifica matemáticamente la direccionalidad de las explicaciones locales: los valores positivos aumentan el *log-odds* empujando la probabilidad hacia la clase de riesgo (Clase 1), mientras que los valores negativos lo reducen, actuando como factores protectores hacia la salud (Clase 0).

6.3.2.2. Justificación Matemática de las Contribuciones en EBM

En el contexto de la clasificación binaria (como la predicción de complicaciones renales), la función de enlace g utilizada en el modelo EBM es la función *logit*. Por lo tanto, el modelo opera internamente en el espacio de *log-odds* (logaritmo de la razón de probabilidades). La suma del intercepto (β_0) y las contribuciones individuales de cada variable determina el *log-odds* final de que un paciente pertenezca a la clase positiva (riesgo de enfermedad):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j) \tag{6.2}$$

Donde p es la probabilidad de que la instancia pertenezca a la clase de riesgo (Clase 1).

Esta naturaleza aditiva en el espacio *logit* es la que justifica matemáticamente la direccionalidad y el color de las barras en las explicaciones locales del modelo:

- Valores positivos (barras naranjas): cuando la función de forma da como resultado $f_j(x_j) > 0$, la característica aumenta el valor del *log-odds*. Matemáticamente, esto empuja la probabilidad p hacia 1, lo que se traduce clínicamente en un aumento directo del riesgo estimado de desarrollar la complicación.

- Valores negativos (barras azules): cuando $f_j(x_j) < 0$, la característica reduce el valor del *log-odds*. Esto empuja la probabilidad p hacia 0, actuando clínicamente como un factor protector o un indicador de salud, desplazando la predicción final hacia la ausencia de complicaciones (Clase 0).

Para obtener la probabilidad final que emite el modelo, el *log-odds* resultante se transforma de regreso al espacio de probabilidad $[0, 1]$ mediante la función logística inversa:

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j))}} \tag{6.3}$$

De esta manera, la magnitud y el signo de cada contribución en EBM representan cambios aditivos exactos en el logaritmo de la probabilidad, demostrando de forma transparente qué variables suman riesgo y cuáles actúan como factores protectores para cada paciente en particular.

6.3.2.3. SHapley Additive exPlanations (SHAP)

Para analizar la interpretabilidad a nivel de instancia (predicciones individuales), se emplea el marco de trabajo SHAP. Este método se basa en la teoría de juegos cooperativos para asignar de manera justa la contribución de cada variable clínica al resultado de la predicción final.

SHAP aproxima la predicción del modelo mediante un modelo de explicación aditivo, definido matemáticamente como:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{6.4}$$

Donde:

- g es el modelo explicativo.
- $z' \in \{0, 1\}^M$ es el vector de características simplificadas (indicando si una característica está presente o ausente).
- M es el número de características de entrada.
- ϕ_0 es el valor base (*base value*), el cual representa la salida esperada del modelo sobre todo el conjunto de entrenamiento ($E[f(x)]$).
- ϕ_j es la contribución individual de la característica j (el valor de Shapley).

Los valores de Shapley (ϕ_j) se calculan mediante la siguiente ecuación:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_x(S \cup \{j\}) - f_x(S) \right] \tag{6.5}$$

Donde:

- F es el conjunto de todas las características.
- S es un subconjunto de características que no contiene a la característica j.
- $f_x(S)$ es la predicción del modelo condicionado a los valores de las características en S.
- La diferencia $f_x(S \cup \{j\}) - f_x(S)$ representa el impacto marginal de añadir la característica j al subconjunto S.

Esta formulación matemática establece que la predicción final para un paciente específico, $f(x)$, es estrictamente la suma del riesgo base promedio más la contribución marginal de cada característica y sus interacciones:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{6.6}$$

6.4. Métricas de evaluación

La matriz de confusión (Erbani et al., 2024) es una herramienta que resume los conteos de predicciones correctas e incorrectas. En ella se representan los verdaderos positivos (TP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (TN). Esta matriz permite visualizar el rendimiento del modelo y sirve como base para calcular métricas derivadas como exactitud, precisión y sensibilidad.

La exactitud (*Accuracy*) se refiere a la proporción de predicciones correctas, tanto positivas como negativas, respecto al total de observaciones (Mengle y Gurmendez, 2019). La cual se calcula en la Ec. (6.7) y es una medida global del desempeño del modelo, aunque puede resultar poco informativa cuando los datos están desbalanceados.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{6.7}$$

La precisión (*Precision*) mide la proporción de instancias clasificadas como positivas que realmente lo son (Cohen et al., 2024). La cual se define en la Ec. (6.8), y es útil para minimizar falsos positivos; resulta importante en contextos donde una clasificación errónea positiva tiene un costo elevado.

$$Precision = \frac{TP}{TP + FP} \tag{6.8}$$

La sensibilidad (*Recall*) (Cohen et al., 2024) indica la proporción de positivos reales que fueron correctamente identificados por el modelo. Se calcula con la Ec. (6.9) y es crucial en situaciones donde no detectar un positivo es más costoso, como en diagnósticos médicos.

$$Recall = \frac{TP}{TP + FN} \tag{6.9}$$

6.5. Trabajos relacionados

En esta sección se analiza la literatura que fundamenta la propuesta metodológica del presente estudio. De un total de 24 artículos revisados, la Tabla 6.2 sintetiza los 11 trabajos más representativos en la predicción de enfermedad renal crónica y diabetes. La selección priorizó investigaciones validadas sobre los conjuntos de datos *UCI Chronic Kidney Disease* y *DiabetIA*, así como estudios que integran técnicas de Inteligencia Artificial Explicable (XAI).

En el ámbito de la medicina computacional, la adopción de modelos predictivos ha estado tradicionalmente condicionada por un compromiso entre precisión y transparencia. Si bien los modelos de aprendizaje automático de alto rendimiento han demostrado capacidad para detectar patrones complejos en datos clínicos, su naturaleza opaca limita su incorporación en sistemas de soporte a la decisión médica, donde la interpretabilidad constituye un requisito esencial para la validación y la confianza clínica.

De manera complementaria, la literatura reciente muestra una consolidación de herramientas interpretables como SHAP y los modelos aditivos generalizados explicables, los cuales permiten descomponer la contribución individual de cada variable en modelos complejos o, alternativamente, construir modelos intrínsecamente interpretables con capacidad para capturar relaciones no

lineales sin sacrificar claridad estructural. Esta tendencia metodológica ha sido aplicada tanto en modelos de ensamble de alto rendimiento como en arquitecturas tabulares modernas, evidenciando una transición progresiva hacia enfoques que equilibran desempeño predictivo y comprensión clínica.

Bajo esta perspectiva, la selección de estudios incluidos en la Tabla 6.2 prioriza investigaciones que no solo reportan métricas de desempeño, sino que incorporan explícitamente mecanismos de interpretabilidad, con el objetivo de analizar la evolución metodológica desde enfoques probabilísticos tradicionales hasta estrategias integradoras basadas en XAI.

La revisión sistematizada presentada en la Tabla 6.2 evidencia una evolución progresiva en los enfoques aplicados a la predicción de enfermedades renales y metabólicas. Los primeros trabajos se centran en modelos probabilísticos y clasificadores tradicionales, destacando el uso de redes bayesianas y métodos basados en características fisiológicas. Posteriormente, se observa una transición hacia modelos de ensamble de alto rendimiento como XGBoost y Random Forest, optimizados mediante técnicas de selección de características híbridas.

En el contexto nacional, estudios recientes han explorado arquitecturas basadas en atención para la identificación de nefropatía diabética en población mexicana, empleando modelos Transformer aplicados a datos tabulares clínicos. Estos trabajos evidencian el potencial de enfoques de Aprendizaje profundo en el conjunto de datos DiabetIA, aunque persiste la necesidad de integrar mecanismos explícitos de interpretabilidad estructural que faciliten la validación clínica de los hallazgos.

Más recientemente, el énfasis se ha desplazado hacia la interpretabilidad, ya sea mediante modelos intrínsecamente explicables como Explainable Boosting Machines (EBM) o a través de técnicas post-hoc como SHAP aplicadas a modelos de alto desempeño. No obstante, la literatura muestra una brecha entre desempeño predictivo, interpretabilidad estructural y validación clínica en contextos reales de enfermedad renal asociada a diabetes.

En este sentido, el presente estudio propone una combinación entre modelos intrínsecamente interpretables (EBM) con técnicas explicativas post-hoc (SHAP), evaluados bajo un análisis comparativo con métodos de ensamble tradicionales, con el objetivo de equilibrar precisión diagnóstica y transparencia clínica en la detección temprana de insuficiencia renal.

6.5.1. Configuración experimental

Tras la etapa de limpieza y depuración, las variables numéricas fueron estandarizadas mediante `StandardScaler`, garantizando una media $\mu = 0$ y desviación estándar $\sigma = 1$. Posteriormente, el conjunto de datos fue dividido utilizando muestreo estratificado en una proporción 80 % para entrenamiento y 20 % para prueba, preservando la distribución de la variable objetivo. Con el fin de asegurar la reproducibilidad del estudio, se fijó una semilla aleatoria global (`random_state = 0`).

El clasificador Gaussian Naïve Bayes se implementó bajo el supuesto de normalidad de las variables predictoras. El Árbol de decisión se entrenó utilizando el criterio de impureza Gini, sin restricción en la profundidad máxima, permitiendo el crecimiento completo del modelo. Para la Support Vector Machine (SVM), se empleó un kernel radial (RBF) con un parámetro de regularización $C = 1.0$. Este valor establece un equilibrio óptimo entre la maximización del margen de decisión y la tolerancia a errores de clasificación durante el entrenamiento. De este modo, se restringe la complejidad de la frontera de decisión para mitigar el riesgo de sobreajuste (*overfitting*), garantizando que el modelo mantenga una alta capacidad de generalización al predecir sobre nuevos casos clínicos.

El modelo de Gradient boosting se configuró con 100 estimadores y una tasa de aprendizaje de 0.1. De manera similar, el Random forest se entrenó con 100 árboles y criterio Gini para la división de nodos. Finalmente, el Explainable Boosting Machine (EBM) incorporó 10 interacciones automáticas y 8 *outer bags* para validación interna, conforme a la parametrización estándar del algoritmo.

Para el análisis de explicabilidad post-hoc, se aplicó `SHAP TreeExplainer` al modelo Random forest, permitiendo el cálculo eficiente de valores de Shapley a partir de la estructura de los árboles y la generación de explicaciones locales de las predicciones individuales.

6.6. Resultados y discusión

6.6.1. Algoritmos de Aprendizaje automático

En la Tabla 6.3 se presentan los resultados experimentales correspondientes a los algoritmos tradicionales de aprendizaje automático y al modelo explicable seleccionado considerando las métricas de *Accuracy* (exactitud), *Precision* (precisión) y *Recall* (sensibilidad), con el propósito de evaluar las diferencias en rendimiento e interpretabilidad. Posteriormente, en la Tabla 4 se muestran las matrices de confusión correspondientes, las cuales permiten analizar con mayor detalle los aciertos y errores de clasificación de cada modelo.

Los resultados de las tablas anteriores muestran un rendimiento competitivo entre los modelos evaluados. Los algoritmos de ensamble (Random forest, Gradient boosting y EBM) superaron a los modelos simples con una exactitud promedio de 0.78, demostrando mayor estabilidad en la clasificación de complicaciones renales. Destaca el modelo SVM, que alcanzó la sensibilidad más alta (0.88), cualidad crítica en el diagnóstico médico para minimizar la omisión de casos positivos (falsos negativos).

El valor diferenciador del Explainable Boosting Machines (EBM) radica en que, manteniendo una competitividad técnica similar a los modelos de caja negra como Random forest, ofrece una transparencia intrínseca. El análisis de las matrices de confusión confirma que el EBM logra un equilibrio óptimo entre precisión y sensibilidad, sin sesgos significativos. Esta capacidad de interpretación nativa facilita la validación clínica y establece una base sólida para el análisis profundo mediante SHAP, permitiendo identificar las causas subyacentes de cada predicción en el contexto de la nefropatía diabética.

6.6.2. Correlación lineal

Para evaluar la relación lineal entre las características clínicas y la presencia de complicaciones renales (variable objetivo `e112`), se llevó a cabo un análisis de correlación de Pearson. La magnitud de la asociación se interpretó de acuerdo con el valor absoluto del coeficiente de correlación (r), clasificándose como despreciable ($0.00-0.10$), débil ($0.10-0.39$), moderada ($0.40-0.69$), fuerte ($0.70-0.89$) y muy fuerte ($0.90-1.00$) (Schober et al., 2018). Los resultados obtenidos se

muestran en la Figura 6.3 y se sintetizan en la Tabla 6.5.

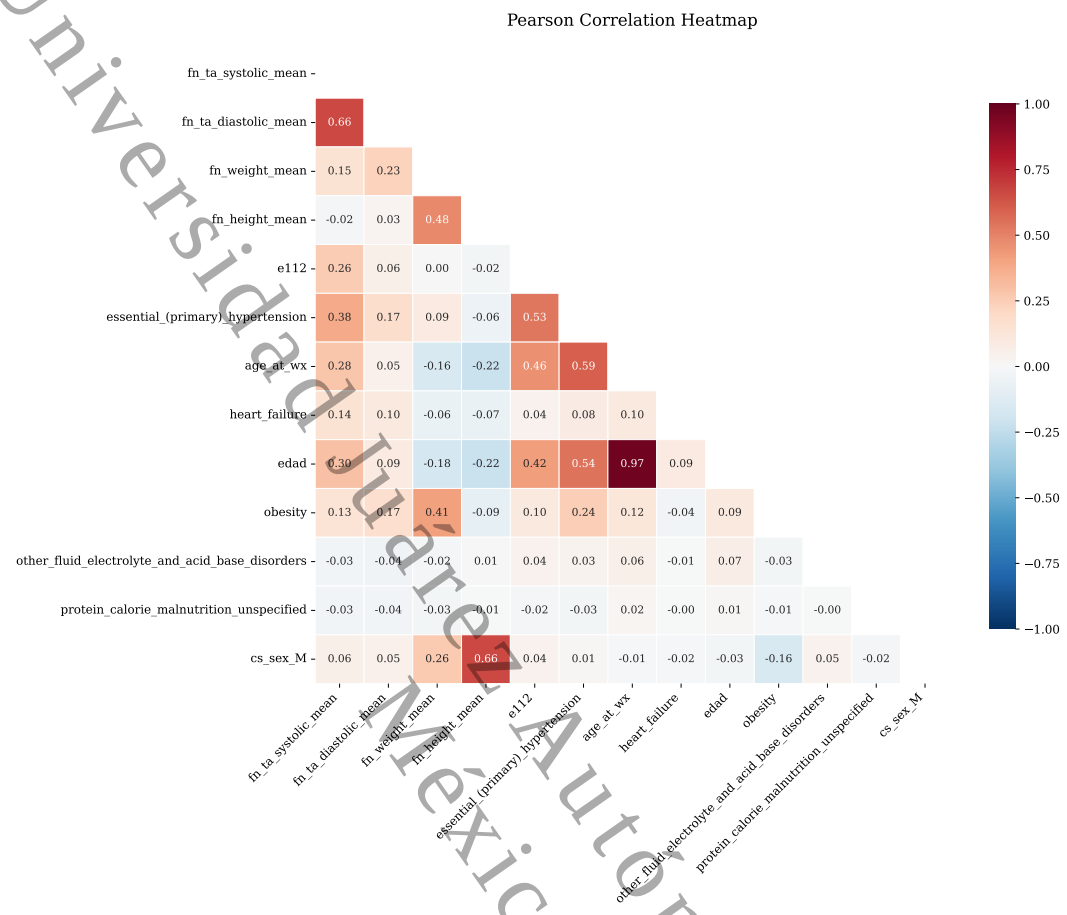

Figura 6.3. Mapa de Calor de Correlaciones de Pearson (Triángulo Inferior.

Los coeficientes obtenidos revelan que la hipertensión esencial primaria exhibe la correlación moderada más alta con la variable `e112` ($r = 0.53$), hallazgo consistente con la literatura médica que identifica a la hipertensión como un factor primario en la progresión de la nefropatía diabética. De manera similar, las variables temporales y de envejecimiento, como la edad al diagnóstico ($r = 0.46$) y la edad actual ($r = 0.42$), presentan correlaciones moderadas, lo cual sugiere que la cronicidad de la diabetes incrementa el riesgo de afectación renal.

La presión sistólica media muestra una correlación débil positiva ($r = 0.26$), mientras que factores como la obesidad ($r = 0.10$), la insuficiencia cardíaca ($r = 0.04$) y el sexo masculino ($r = 0.04$) presentan asociaciones despreciables con `e112`. Destaca también una correlación moderada entre `age_at_wx` y `edad` ($r = 0.97$), lo que indica una alta colinealidad entre ambas variables y deberá considerarse en etapas posteriores de modelado.

Es importante señalar que, si bien algunos coeficientes parecen bajos, su relevancia clínica puede ser significativa en contextos no lineales. Esto justifica la implementación de modelos no lineales y técnicas de IA Explicable (EBM y SHAP) en las secciones posteriores, dado que permiten capturar interacciones complejas que un análisis de correlación lineal simple podría pasar por alto.

6.6.3. Analisís de algoritmos interpretables

6.6.3.1. Analisís global de EBM

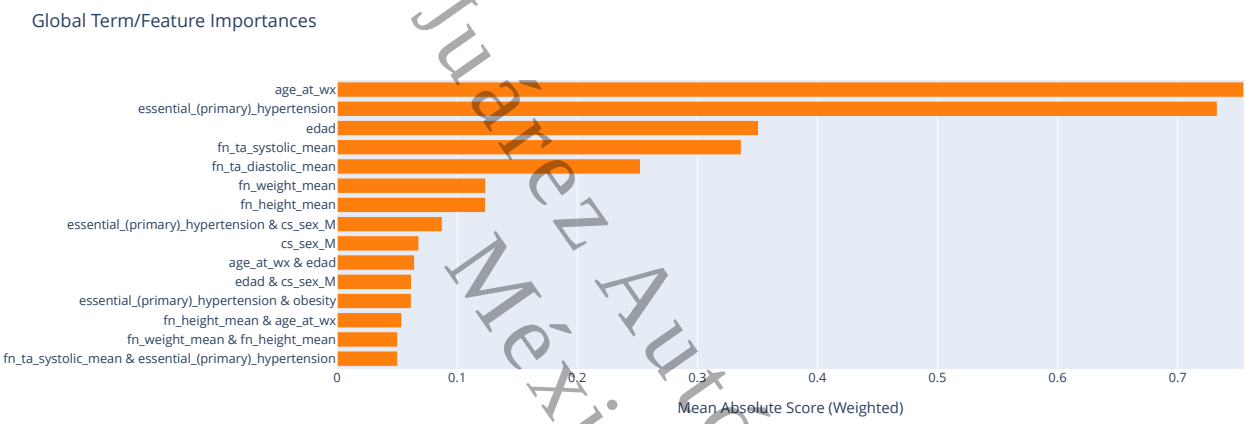

Figura 6.4. Importancia global de características del modelo EBM. El gráfico clasifica las variables según su puntuación media absoluta (ponderada), evidenciando la predominancia de la edad al diagnósticoy la hipertensión arterial como los principales predictores de nefropatía.

El análisis de las importancia globales de características (Figura 6.4), inherente al modelo EBM, revela una jerarquía predictiva. Los factores más determinantes son la edad al diagnóstico (`age_at_wx`) y la hipertensión esencial, seguidos por la edad actual. Asimismo, se observa que las interacciones entre variables (como peso/estatura con edad) enriquecen significativamente el modelo, confirmando que la progresión de la enfermedad depende de múltiples dimensiones no lineales que modelos simples pasarían por alto.

6.6.3.2. Análisis de decisiones con SHAP (*Decision plot*)

El gráfico de decisión de SHAP (Figura 6.5), corrobora la jerarquía de importancia global previamente identificada por el modelo EBM. En este gráfico, el eje vertical enlista las características

ordenadas por importancia, mientras que el eje horizontal muestra cómo cada característica desplaza la probabilidad de predicción (valor de salida del modelo) para las distintas observaciones.

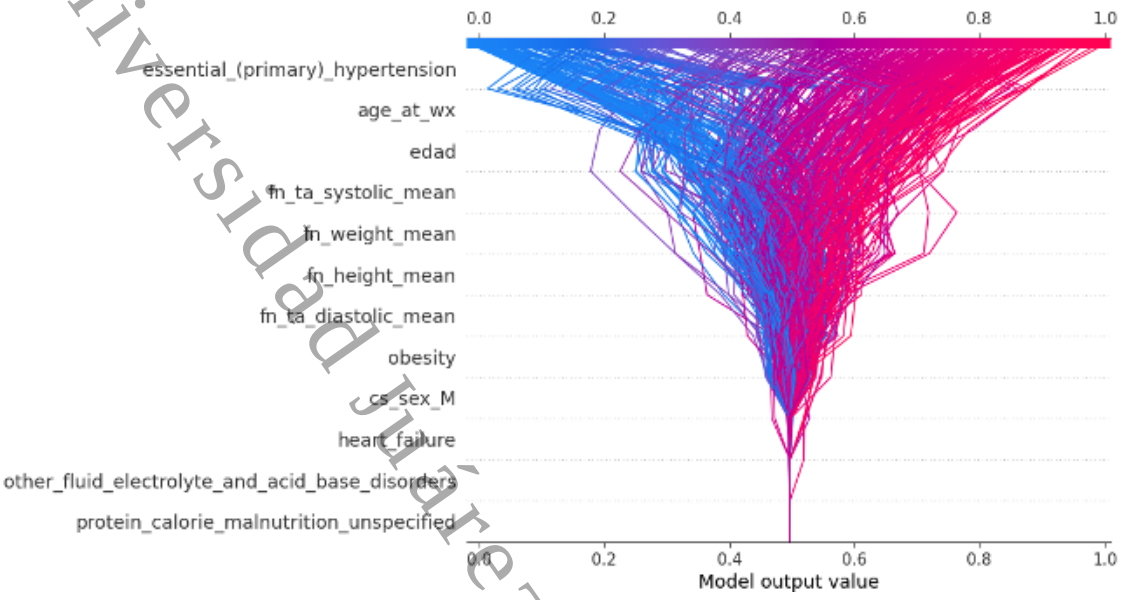

Figura 6.5. SHAP Decision Plot. Las líneas muestran cómo cada característica contribuye a mover el valor de salida del modelo desde la base hasta la predicción final. El color rojo indica valores altos de la característica y el azul valores bajos. Se observa cómo la hipertensión y la edad (arriba) son determinantes para desplazar la predicción hacia la derecha (mayor riesgo).

Se puede observar en la Figura 6.5 que las características ubicadas en la parte superior, como la hipertensión (`essential_primary_hypertension`) y las variables relacionadas con la edad, se confirman como los predictores más influyentes. Esto genera una alta confianza en la robustez del modelo al alinear los hallazgos de dos metodologías distintas (EBM y SHAP). La importancia de indicadores como la presión arterial sistólica media y la diastólica media, así como las métricas de peso y altura, también se mantiene consistente, apareciendo en el tercio superior del gráfico.

La principal ventaja de este gráfico sobre la Importancia Global del EBM es la visualización de la dirección y magnitud del impacto acumulativo:

- Las líneas de color rojo (que indican presencia de hipertensión o valores altos) muestran un desplazamiento marcado hacia el extremo derecho del eje x (hacia 1.0). Esto indica que la hipertensión es el factor individual que más empuja la predicción hacia la probabilidad de insuficiencia renal.
- De manera similar, las trayectorias correspondientes a pacientes de mayor edad (color rojo)

tienden a moverse hacia la derecha, confirmando que el envejecimiento contribuye positivamente al riesgo calculado.

- Las variables en la parte inferior, como `heart_failure` o desnutrición (`protein_calorie_malnutrition_unspecified`), muestran líneas más verticales y centradas, indicando un impacto neutro o marginal en la mayoría de los casos.

Mientras que EBM proporcionó una importancia global robusta, este análisis post-hoc con SHAP valida la jerarquía de predictores y revela la dinámica de las decisiones. Confirma que la hipertensión esencial y la edad no solo son importantes, sino que sus valores altos son los que ejercen la mayor contribución positiva (aumento de riesgo) en la salida del modelo. La convergencia entre la interpretabilidad intrínseca del EBM y el análisis de SHAP refuerza la confianza clínica, garantizando que los efectos de las variables son consistentes y validables.

6.6.4. Analisís local de EBM

6.6.5. Interpretabilidad local (análisis de casos específicos)

En esta sección se realiza un análisis de interpretabilidad local con el fin de examinar la influencia individual de cada característica en las predicciones del modelo. El objetivo es identificar cuáles variables tienen un mayor impacto en la clasificación de un paciente específico, permitiendo comprender las razones detrás de cada decisión.

Para ejemplificar este enfoque, se analizan casos representativos: predicciones correctas de enfermedad y salud (verdaderos positivos y negativos), así como casos de error, lo que permite observar cómo las interacciones y los valores de las variables influyen en la clasificación final.

6.6.5.1. Caso 1: predicción correcta de salud (TN)

El modelo EBM predijo correctamente que el paciente no presenta complicaciones renales (Clase 0), con una probabilidad de 0.679, coincidiendo con la etiqueta real. Con base en la justificación matemática del algoritmo (espacio *log-odds*), la Figura 6.6 muestra el desglose de contribuciones locales: las barras azules ($f(x) < 0$) reducen la probabilidad desplazando la predicción

hacia la salud, mientras que las naranjas ($f(x) > 0$) aportan peso matemático aumentando el riesgo estimado.

Al aplicar la función logística inversa a la suma del intercepto (β_0) y estas contribuciones, se obtiene la probabilidad final que consolida la clasificación del paciente como sano.

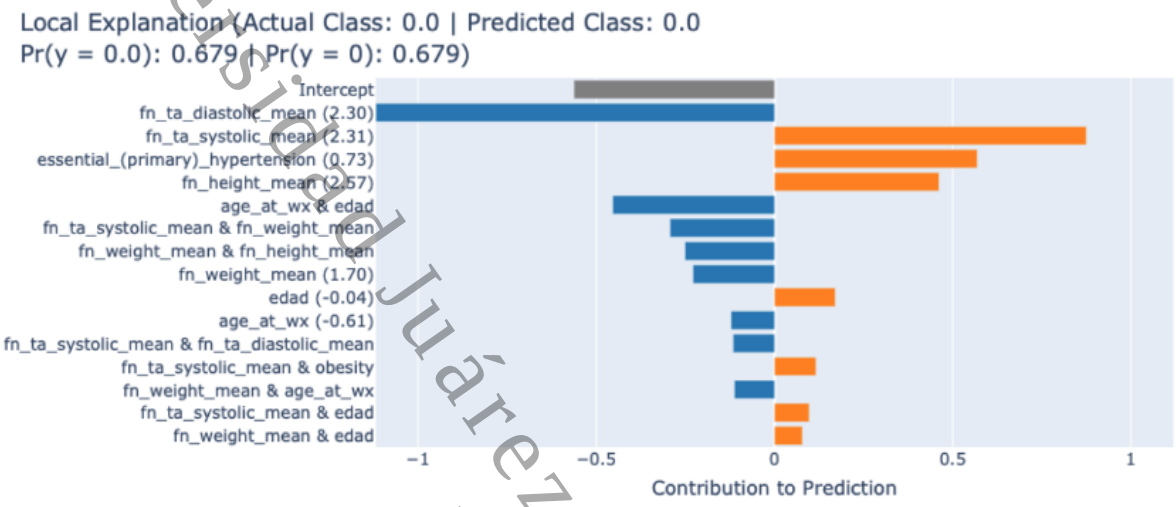

Figura 6.6. Explicación Local del Caso 1 (TN). El modelo clasifica correctamente al paciente como sano. Se observa cómo la presión diastólica (barra azul superior) y las interacciones de edad actúan como factores protectores que contrarrestan algebraicamente los factores de riesgo (barras naranja).

Las características determinantes para esta clasificación fueron:

- Presión Diastólica (`fn_ta_diastolic_mean = 2.30`): Es el factor individual más determinante. Su contribución es fuertemente negativa (barra azul de mayor longitud), actuando como el principal "protector" contra la predicción de riesgo dentro de la suma aditiva.
- Edad al diagnóstico y edad actual: La variable combinada (`age_at_wx & edad`) actúa como un mitigador significativo, reduciendo el valor del *log-odds* y desplazando la predicción hacia la salud renal.
- Peso (`fn_weight_mean = 1.70`): A diferencia de la estatura, el peso en este paciente específico contribuye negativamente al riesgo (barra azul), ayudando a anclar la predicción en la clase sana.

Es interesante notar que no todos los factores apuntan a la salud. La `estatura` (`fn_height_mean`) y la presión sistólica (`fn_ta_systolic_mean`) aparecen como barras naranja de mayor tamaño, aportando peso hacia el riesgo. Sin embargo, el modelo determina que la contribución neta de

la estabilidad diastólica y las interacciones protectoras (como peso-estatura y peso-presión) son suficientes para contrarrestar matemáticamente estos riesgos, resultando en una predicción correcta de ausencia de enfermedad.

6.6.5.2. Caso 2: predicción correcta de enfermedad (TP)

Esta instancia corresponde a un verdadero positivo (TP), donde el modelo clasifica correctamente al paciente como enfermo (Clase 1), con una probabilidad predicha de 0.768. La Figura 6.7 ilustra cómo la acumulación de factores de riesgo supera a los factores protectores.

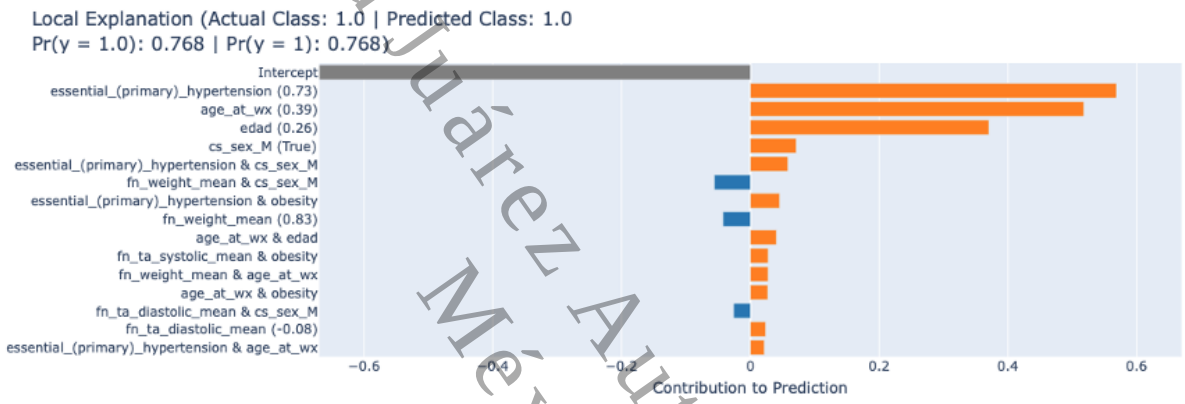

Figura 6.7. Explicación local del caso 2 (TP). A pesar de un intercepto negativo (barra gris) y factores mitigantes como el peso (barras azules), la fuerte influencia de la hipertensión y la edad (barras naranja) empuja la probabilidad final hacia la zona de alto riesgo (0.768).

El desglose de fuerzas revela un conflicto directo entre factores clínicos. La predicción es impulsada dominantemente por tres factores (barras naranja) que ejercen un fuerte efecto acumulativo:

- La hipertensión esencial (`essential_primary_hypertension`) es el motor principal del riesgo.
- La edad al diagnóstico (`age_at_wx`) y la edad actual (`edad`) suman, confirmando la preponderancia de los factores demográficos y la cronicidad en el desarrollo de la complicación renal.
- A pesar de la robusta predicción de riesgo, Se destaca que el peso medio favorable (`fn_weight_mean`) y su interacción con el sexo masculino (`cs_sex_M`) actúan como mitigadores (barras azules), reduciendo parcialmente la probabilidad de riesgo.

- Un hallazgo importante es que el valor base del modelo se sitúa en el lado negativo (barra gris). Esto implica que la tendencia base poblacional asume un riesgo bajo; sin embargo, en este caso específico, la magnitud del riesgo aportado por la hipertensión y la edad es tan severa que logra revertir la tendencia base, empujando la predicción final hasta el 76.8 %.

De esta manera, el modelo EBM demuestra su capacidad para cuantificar exactamente cómo los principales factores cardiovasculares logran superar el efecto protector del peso, resultando en una clasificación final de alto riesgo que es clínicamente validable.

6.6.5.3. Caso 3: análisis de error (FP)

En este escenario, el modelo EBM presenta una discrepancia con la etiqueta real, clasificando erróneamente al paciente en la categoría de riesgo (Clase 1) con una probabilidad muy alta de **0.874**, cuando en realidad pertenece a la clase sana (Clase 0). La Figura 6.8 permite entender la anatomía de esta “falsa alarma”.

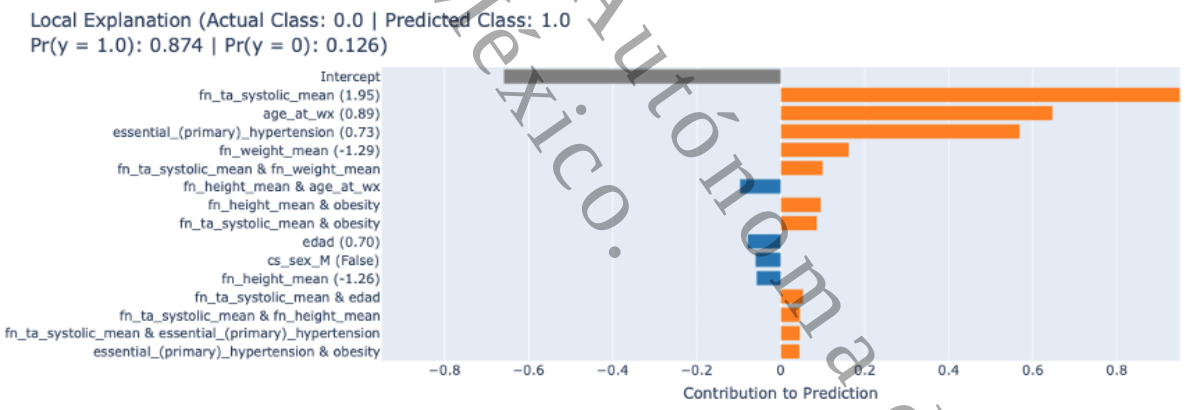

Figura 6.8. Explicación Local del Caso 3 (FP). El modelo sobreestima el riesgo debido a una presión sistólica elevada (barra naranja dominante), ignorando las señales de salud (barras azules).

Al observar las contribuciones locales, se identifica por qué el modelo falló:

- La presión sistólica media (`fn_ta_systolic_mean`) actúa como el principal impulsor de la clasificación errónea. Con una contribución de 1.95, esta sola variable domina la lógica del modelo.
- La edad al diagnóstico (`age_at_wx = 0.89`) y la hipertensión (`0.73`) refuerzan la señal de alarma. El modelo interpreta este perfil como “vulnerable” , asumiendo erróneamente que la

presión alta implica daño renal.

- Variables que intentan desplazar la predicción hacia la salud (barras azules), como la estatura (`fn_height_mean`) y la edad actual (`edad`), presentan una magnitud muy reducida. No tienen la fuerza suficiente para contrarrestar el peso de la presión sistólica.

Este análisis sugiere que el modelo es altamente sensible a la presión sistólica. Clínicamente, esto genera una falso positivo en pacientes que quizás tienen hipertensión aislada pero riñones sanos. Sin embargo, en un contexto de tamizaje médico, este sesgo conservador (preferir un falso positivo a omitir un caso real) suele ser aceptable.

6.6.5.4. Caso 4: error crítico (FN)

Este caso ilustra el error más sensible en medicina: un fallo de sensibilidad (FN). El modelo clasificó a un paciente con complicación renal real como sano (Clase 0), con una probabilidad de 0.637 a favor de la salud. La Figura 6.9 expone la causa de esta omisión.

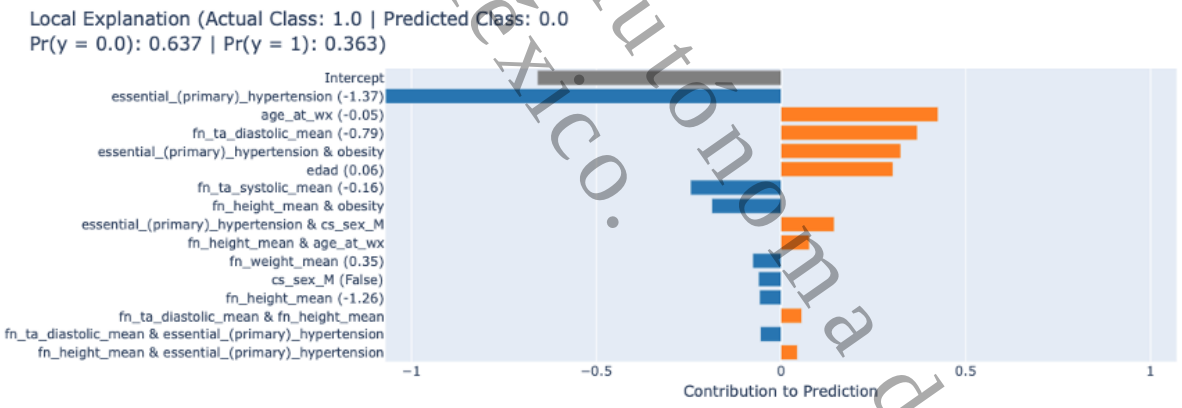

Figura 6.9. Explicación local del caso 4 (FN). La ausencia de hipertensión (barra azul masiva) neutraliza las señales de alerta (barras naranjas) que sí detectaron otros indicadores como la edad y la presión diastólica.

El desglose de fuerzas muestra cómo una sola variable puede sesgar la decisión:

- La variable `essential_primary_hypertension` presenta una contribución negativa masiva de -1.37 (barra azul dominante). Al no detectar hipertensión marcada, el modelo asume erróneamente un diagnóstico de salud, restando una cantidad inmensa de puntaje al riesgo acumulado.

- A pesar del error final, el algoritmo sí detectó señales de peligro (barras naranjas) que intentaron advertir la enfermedad, como edad al diagnóstico (`age_at_wx`) y presión diastólica (`fn_ta_diastolic_mean`). Ambas generan barras naranjas considerables, indicando que por edad y perfil diastólico el paciente sí tenía riesgo.
- La combinación de hipertensión y obesidad también sumó puntos al riesgo.

La magnitud de la contribución negativa de la hipertensión es tanta que prácticamente anula a los indicadores de riesgo válidos. Esto sugiere que el modelo podría estar dependiendo excesivamente de la hipertensión como criterio de descarte, lo cual es una oportunidad clara para re-calibrar el peso de esta variable en futuras iteraciones.

6.6.6. Análisis post-hoc con SHAP y Random forest

En esta sección se realiza de manera análoga al análisis efectuado con el modelo EBM, un estudio de interpretabilidad local con el propósito de examinar la influencia individual de cada característica en las predicciones del modelo. El objetivo es identificar cuáles variables ejercen un mayor impacto en la clasificación de cada paciente a través del método SHAP.

A diferencia del EBM, SHAP se aplica aquí sobre un modelo de tipo caja negra, en este caso Random forest, con el fin de obtener una explicación detallada de cómo cada variable contribuye positiva o negativamente a la predicción final. De forma similar al procedimiento previo, se analizan casos representativos para comparar la coherencia de ambos enfoques interpretativos.

6.6.6.1. Caso 1: predicción correcta de salud (TN)

Para validar la interpretación del modelo Random forest a nivel de instancia, se empleó un diagrama de fuerza (*force plot*), el cual permite visualizar la competencia entre las características clínicas y materializar la ecuación aditiva de SHAP (Ecuación 6.6).

En este caso de verdadero negativo (TN), como se muestra en la Figura 6.10, el valor base del modelo (*base value*), que representa la expectativa promedio del conjunto de datos, se sitúa en $\phi_0 = 0.4982$. La suma de las contribuciones marginales de cada variable ($\sum_{j=1}^{M} \phi_j$) genera un impacto neto negativo que logra desplazar la predicción final hacia la izquierda, alcanzando un

valor exacto de $f(x) = 0.26$. Este descenso respecto al riesgo base consolida la clasificación del paciente como sano.

Figura 6.10. SHAP force plot para el caso 1 (TN). Las líneas azules (factores de protección con $\phi_j < 0$), guiadas por la ausencia de hipertensión y la edad, empujan la predicción desde el valor base ($\phi_0 = 0.4982$) hacia un riesgo bajo ($f(x) = 0.26$), contrarrestando las líneas rojas de menor fuerza.

Las variables que ejercen la mayor influencia para disminuir la probabilidad de enfermedad (barras azules, donde $\phi_j < 0$) son:

- Ausencia de hipertensión (`essential_primary_hypertension = 0`): Es el factor predominante que empuja la predicción hacia la izquierda, restando valor a la suma total y alejando al paciente del umbral de riesgo.
- Edad (`age_at_wx = 42` y `edad = 49`): Los factores etarios actúan como protectores en este perfil relativamente joven, contribuyendo significativamente a la reducción de la predicción final.
- La estatura (`fn_height_mean = 1.49`) y la presión sistólica controlada (`110`) aportan valores de Shapley negativos que refuerzan el diagnóstico de salud renal.

Por el contrario, variables como la presión diastólica (`fn_ta_diastolic_mean = 75`) y el peso (83) ejercen una presión aditiva hacia el riesgo (barras rojas, donde $\phi_j > 0$). Sin embargo, su impacto marginal es visual y matemáticamente menor, por lo que no logran contrarrestar la suma de las fuerzas de protección. Este balance demuestra que el modelo pondera correctamente la estabilidad hemodinámica y la ausencia de comorbilidades crónicas.

6.6.6.2. Caso 2: predicción correcta de enfermedad (TP)

En este escenario de verdadero positivo (TP), el modelo Random forest predijo correctamente la presencia de complicaciones renales (Clase 1), partiendo de un valor base de ϕ_0= 0.4982. El

gráfico de fuerza (Figura 6.11) permite observar cómo los factores de riesgo superan ampliamente a las variables de protección, desplazando la probabilidad final hasta $f(x) = 0.72$.

Figura 6.11. SHAP force plot para el caso 2 (TP). Las fuerzas rojas (factores de riesgo con $\phi_j > 0$), impulsadas dominantemente por la presencia de hipertensión y la edad avanzada, elevan la predicción desde el valor base ($\phi_0 = 0.4982$) hasta una probabilidad alta de enfermedad ($f(x) = 0.72$), superando a las fuerzas azules de mitigación.

Las variables que ejercen la mayor influencia positiva hacia la clase de enfermedad (barras rojas) son:

- Hipertensión (`essential_primary_hypertension = 1`): Se identifica como el motor principal de la predicción. La presencia de esta comorbilidad ejerce la mayor presión visual para clasificar al paciente en el grupo de riesgo.
- La edad avanzada actual `edad = 87` y al momento del diagnóstico `age_at_wx = 74` actúan como factores críticos acumulativos que incrementan sustancialmente la probabilidad de insuficiencia renal.
- El elevado peso (`fn_weight_mean = 84.65`) y la presión diastólica (`73.75`), aunque en rangos aparentemente controlados, suman carga adicional al riesgo total en este perfil específico.
- Por el contrario, solo unos pocos factores intentan reducir la probabilidad (barras azules). La estatura (`fn_height_mean = 1.77`) y la obesidad (`obesity = 1`) actúan como fuerzas menores de oposición. Sin embargo, su impacto es visualmente insignificante frente a la magnitud de la hipertensión y la edad.

Este balance confirma que, para el modelo, la comorbilidad hipertensiva combinada con la longevidad son los determinantes definitivos para diagnosticar la complicación renal en este paciente.

6.6.6.3. Caso 3: análisis de error (FP)

Finalmente, se analiza un caso de falso positivo (FP), donde el modelo clasifica erróneamente a un paciente sano (Clase 0) dentro de la categoría de riesgo (Clase 1). Como se observa en la Figura 6.12, se tiene el valor base ϕ_0=0.4982, donde la predicción se dispara hasta $f(x)$= 0.77, impulsada por una acumulación de factores de riesgo que el modelo sobrevalora.

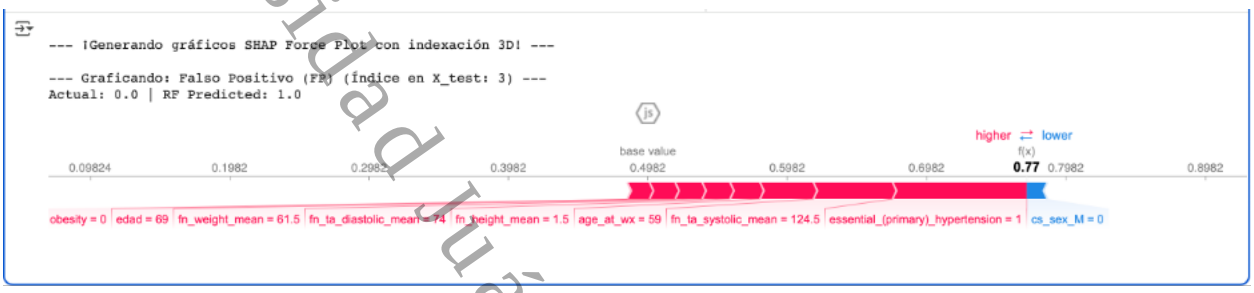

Figura 6.12. SHAP force plot para el caso 4 (FP). Las fuerzas rojas (factores de riesgo con $\phi_j > 0$), impulsadas por la presencia de hipertensión y la edad avanzada, generan una "falsa alarma". Esta acumulación eleva la predicción desde el valor base ($\phi_0 = 0.4982$) hasta un alto riesgo estimado ($f(x) = 0.77$), llevando al modelo a clasificar erróneamente a un paciente sano.

Las variables que provocan este falso diagnóstico (barras rojas) son:

- Hipertensión (`essential_primary_hypertension = 1`): actúa nuevamente como el predictor con mayor carga. El modelo asume que la mera presencia de hipertensión implica daño renal inminente.
- Presión sistólica elevada (`124.5`) y edad (`69`):refuerzan la lógica del modelo sobre la vulnerabilidad del paciente.

La única fuerza de oposición es el sexo femenino (`cs_sex_M = 0`), cuya magnitud (barra azul) es muy pequeña frente a la combinación de hipertensión y edad. Clínicamente, esto representa un sesgo hacia la sensibilidad: el sistema prefiere alertar un posible riesgo (aunque sea falso) antes que omitir una enfermedad.

6.6.6.4. Caso 4: error crítico (FN)

En este escenario de falso negativo (FN), el modelo Random forest falla al clasificar a un paciente con complicación renal real (Clase 1) como sano (Clase 0). A pesar de partir de un valor

base de ϕ_0=0.4982, la influencia acumulada de las variables de protección (barras azules) mueve la predicción hasta $f(x)$ = 0.38, por debajo del umbral de decisión.

Figura 6.13. SHAP force plot del caso 3 (FN). Las fuerzas azules masivas (factores con $\phi_j < 0$), impulsadas dominantemente por la ausencia de hipertensión, actúan como el principal distractor. Estas fuerzas desplazan la predicción desde el valor base ($\phi_0 = 0.4982$) hacia la zona de bajo riesgo, contrarrestando erróneamente la influencia de los indicadores de riesgo reales (fuerzas rojas menores) y resultando en una clasificación incorrecta de salud.

Las variables que ejercen la mayor influencia para "ocultar' el riesgo son:

- Ausencia de hipertensión (`essential_primary_hypertension = 0`): Al igual que en el modelo EBC, la ausencia de hipertensión actúa como el distractor principal. Ejerce una presión masiva hacia la izquierda (azul), indicando salud y neutralizando al resto de variables.
- Edad moderada (`edad = 57`): El modelo interpreta este valor como de "bajo riesgo", contribuyendo a la clasificación errónea.
- Presión sistólica normal (`113`): Refuerza la falsa seguridad del algoritmo.

Por el contrario, los factores que intentan advertir sobre la complicación (barras rojas) como la presión diastólica (`72.22`) y el peso (`50.27`), no tienen la magnitud suficiente para revertir la tendencia. Este análisis revela que el modelo depende en exceso de la hipertensión para activar la alerta.

6.6.7. Discusión de la convergencia de interpretabilidad de EBM y SHAP

El análisis comparativo entre la interpretabilidad intrínseca del EBM y la aproximación post-hoc de SHAP revela una convergencia estructural. Ambos modelos coinciden en identificar a la hHipertensión esencial y la edad como los ejes rectores de la predicción.

Sin embargo, el análisis cruzado de los errores expone como vulnerabilidad compartida la sobre-dependencia de la hipertensión:

1. En los falsos negativos (Caso 3) la ausencia de hipertensión actúa como una “barrera”, ocultando otras señales de alerta y provocando que se escapen diagnósticos reales.
2. En los falsos positivos (Caso 4): La sola presencia de hipertensión es suficiente para activar la alarma de enfermedad en pacientes sanos.

Estos hallazgos sugieren la necesidad de recalibrar la ponderación de la variable hipertensión en futuras iteraciones del modelo. Aunque se mantiene como el predictor de mayor contribución, su peso actual tiende a dominar la función de decisión del modelo y puede desplazar señales fisiológicas relevantes en perfiles clínicos atípicos (por ejemplo, pacientes normotensos con daño renal o hipertensos bajo control terapéutico).

6.7. Discusión

La Tabla 6.6 resume la convergencia de hallazgos entre los dos enfoques de explicabilidad utilizados.

En conjunto, los experimentos realizados demuestran que los modelos basados en ensamble y *boosting* alcanzan una precisión diagnóstica competitiva, cercana al 80 %, en la detección de complicaciones renales en la población de estudio.

El análisis de interpretabilidad cruzada mediante SHAP y EBM ratifica que la hipertensión esencial y la edad (avanzada) son los factores con mayor peso predictivo, hallazgo consistente con la evidencia clínica internacional. Sin embargo, la mayor aportación de este estudio radica en la detección de sesgos: ambos modelos tienden a subestimar el riesgo en pacientes normotensos (falsos negativos), un matiz que modelos lineales tradicionales no habrían revelado con tal claridad.

Esta capacidad para desglosar interacciones no lineales posiciona a la Inteligencia Artificial Explicable (XAI) no solo como un mecanismo de predicción, sino como una herramienta de auditoría clínica, vital para implementar una medicina de precisión confiable y transparente en el contexto del sistema de salud mexicano.

6.8. Conclusiones y trabajo futuro

Este estudio ha demostrado la importancia crítica de la Inteligencia Artificial Explicable (XAI) aplicada a la salud pública, utilizando el dataset DiabetIA representativo de la población mexicana. Se logró transformar un diagnóstico binario convencional en una herramienta de prevención efectiva, permitiendo identificar no solo si un paciente presenta riesgo de insuficiencia renal, sino por qué ocurre dicha predicción, desglosando las variables clínicas individuales.

Los principales hallazgos se pueden resumir como sigue:

1. La implementación del Explainable Boosting Machine (EBM) validó que la transparencia del modelo no compromete el rendimiento predictivo. El modelo alcanzó métricas de precisión competitivas frente a algoritmos de "caja negra", estableciendo un mapa de riesgo global donde la hipertensión y la edad se confirmaron como factores dominantes.

2. La aplicación de SHAP sobre modelos complejos (Random forest) funcionó como un mecanismo de auditoría eficaz. No solo corroboró los aciertos del modelo, sino que fue detectó vulnerabilidades diagnósticas, específicamente en los casos de falsos negativos. Se identificó que la ausencia de un diagnóstico explícito de hipertensión puede llevar a los algoritmos a subestimar riesgos latentes, un hallazgo que habría permanecido oculto sin este análisis granular.

3. La integración de EBM y SHAP constituye un marco metodológico robusto para el sistema de salud. Este sistema empodera al personal médico para justificar intervenciones tempranas con evidencia trazable, fomentando la confianza necesaria para adoptar la medicina de precisión en entornos clínicos reales.

A partir de los sesgos detectados en la fase de interpretabilidad, las futuras líneas de investigación deberán centrarse en la recalibración de las variables de alto impacto como la hipertensión. Se propone explorar técnicas de ingeniería de características y remuestreo dirigido que permitan al modelo aumentar su sensibilidad ante perfiles atípicos, como pacientes normotensos con daño renal incipiente, garantizando así una cobertura preventiva más universal.

Referencias del capítulo

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bisong, E. (2019). The Multilayer Perceptron (MLP). En *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 401-405). Apress. https://doi.org/10.1007/978-1-4842-4470-8_31

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth; Brooks/Cole.

Cohen, L., Mansour, Y., Moran, S., & Shao, H. (2024). Probably Approximately Precision and Recall Learning [Disponible en: https://arxiv.org/abs/2411.13029]. *arXiv preprint arXiv:2411.13029*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Dirección General de Epidemiología. (2025). *Informe trimestral del Sistema de Vigilancia Epidemiológica Hospitalaria de Diabetes Mellitus Tipo 2 (SVEHDMT2), tercer trimestre 2025* (inf. téc.). Secretaría de Salud, Gobierno de México. https://www.gob.mx/salud/documentos/diabetes-mellitus-tipo-2-hospitalaria-2025

Erbani, J., Portier, P.-É., Egyed-Zsigmond, E., & Nurbakova, D. (2024). Confusion Matrices: A Unified Theory. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2024.10769075

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer. https://doi.org/10.1007/978-3-319-98074-4

Flores-Custodio, O., & Pancardo-García, P. (2025). A Transformer Network Model for the Diagnosis of Chronic Ischemic Heart Disease in Patients with Type 2 Diabetes Mellitus. *Artificial Intelligence – COMIA 2025*, *2554*, 245-260. https://doi.org/10.1007/978-3-031-97913-2_19

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gaur, L., Biswas, M., Bakshi, S., Gupta, P., Si, T., Mallik, S., & Maulik, U. (2022a). An Integrated Model to Evaluate the Transparency in Predicting Chronic Kidney Disease Using a Trio-Embedded Explainable Model [SSRN Preprint]. https://doi.org/10.2139/ssrn.4129888

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2.ª ed.). O'Reilly Media.

Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, *14*(3687). https://doi.org/10.1038/s41598-024-54375-4

Gogoi, P., & Valan, J. A. (2025). Interpretable Machine Learning for Chronic Kidney Disease Prediction: A SHAP and Genetic Algorithm-Based Approach. *Biomedical Materials & Devices*, *3*, 1384-1402. https://doi.org/10.1007/s44174-024-00262-5

Gómez García, A., López Pineda, A., García Velázquez, L. M., Flores Garrido, M., Álvarez Aguilar, C., & Figueroa Mora, K. M. (2023). Base de Datos DiabetIA [Conjunto de datos]. https://repositorio-salud.conacyt.mx/jspui/handle/1000/296

González, L. R. T. M., & Nolasco, J. A. H. (2025). Modelo Transformer para la identificación de la nefropatía diabética en pacientes Mexicanos. *Research in Computing Science*, *154*(5), 87-101. https://www.rcs.cic.ipn.mx/2025_154_5/

Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. (2024). Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information*, *16*(1), 7. https://doi.org/10.3390/info16010007

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hernández-Gaspar, S. J., Hernández-Ocaña, B., Hernández-Torruco, J., & Chávez-Bosquez, O. (2025). Inteligencia artificial explicable para el análisis de pacientes con insuficiencia renal crónica. *Nova Scientia*, *17*, 1-22. https://doi.org/10.21640/ns.v17iSpecialIa.3630

Huang, Y., Su, S., Chen, J., Zhang, X., Tan, K., & Guo, Q. (2026). An ultrasound-based machine learning model for predicting pelvic adhesions: A SHAP-enhanced XGBoost approach. *DIGITAL HEALTH*, *12*. https://doi.org/10.1177/20552076261416797

IDF. (2021). *International Diabetes Federation Diabetes Atlas* (10th). International Diabetes Federation. https://diabetesatlas.org/

Instituto Nacional de Salud Pública. (2021). *Encuesta Nacional de Salud y Nutrición 2021 sobre Covid-19. Resultados Nacionales* (inf. téc.). INSP. https://ensanut.insp.mx/

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

KDIGO. (2022). *Kidney Disease Improving Global Outcomes - Clinical practice guideline for diabetes management in chronic kidney disease* (inf. téc.). https://kdigo.org/wp-content/uploads/2022/10/KDIGO-2022-Clinical-Practice-Guideline-for-Diabetes-Management-in-CKD.pdf

Khan, M., Ahmed, S., & Rahman, M. (2025). A comparative study of explainable machine learning models with Shapley values for diabetes prediction. *Healthcare Analytics*, *7*, 100390. https://doi.org/10.1016/j.health.2025.100390

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lundberg, S. M., Erion, G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 4768-4777.

Maron, M. E. (1961). Automatic indexing: An experiment with a statistical model. *Journal of the ACM*, *8*(3), 404-417. https://doi.org/10.1145/321075.321084

Mengle, S. S., & Gurmendez, M. (2019). *Mastering Machine Learning on AWS*. Packt Publishing.

Mishra, A. (2021). *Mastering Machine Learning on AWS*. Packt Publishing.

Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book

Nogueira, A. R., Ferreira, C. A., & Gama, J. (2018). Improving acute kidney injury detection with conditional probabilities. *Intelligent Data Analysis*, *22*(6), 1355-1374. https://doi.org/10.3233/IDA-173626

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability [arXiv preprint arXiv:1909.09223]. https://doi.org/10.48550/arXiv.1909.09223

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Raihan, M. J., Khan, M. A., Kee, S. H., & Nahid, A. A. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, *13*(1), 6263. https://doi.org/10.1038/s41598-023-33525-0

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Dataset. https://doi.org/10.24432/C5G020

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Learning*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Swapna, G., Acharya, U. R., VinithaSree, S., & Suri, J. S. (2013). Automated detection of diabetes using higher-order spectral features extracted from heart rate signals. *Intelligent Data Analysis*, *17*(2), 309-326. https://doi.org/10.3233/IDA-130580

Yousefi, L., & Tucker, A. (2022). Identifying latent variables in dynamic Bayesian networks with bootstrapping applied to prediction of type 2 diabetes complications. *Intelligent Data Analysis*, *26*(2), 501-524. https://doi.org/10.3233/IDA-205570

Tabla 6.2. Comparación de trabajos relacionados en enfermedad renal, XAI y el conjunto de datos DiabetIA.

Algoritmos Principales	**Resultados / Enfoque / Contribución**
	Aprendizaje Automático Tradicional y Probabilístico
Redes Bayesianas Nogueira et al., 2018	Uso de probabilidades condicionales para mejorar la detección de lesión renal aguda mediante inferencia clínica estructurada.
Clasificadores basados en VFC Swapna et al., 2013	Detección automatizada de diabetes empleando características espectrales de orden superior extraídas de señales cardiacas, logrando una exactitud del 94 %.
Redes Bayesianas Dinámicas Yousefi y Tucker, 2022	Identificación de variables latentes con bootstrap para la predicción longitudinal de complicaciones de la diabetes tipo 2.
XGBoost Raihan et al., 2023	Optimización de clasificadores para predicción temprana, alcanzando una exactitud del 99.16 %; hemoglobina y albúmina identificadas como predictores principales.
RF, SVM, XGBoost S. K. Ghosh y Khandoker, 2024	Evaluación comparativa de múltiples clasificadores reportando un AUC de 0.96, donde XGBoost superó a los modelos tradicionales.
Algoritmos Genéticos, XGBoost Gogoi y Valan, 2025	Enfoque híbrido de selección de características que mejora significativamente el rendimiento predictivo, alcanzando una exactitud del 99.17 %.
	XAI en la predicción de diabetes
RF, GBM, EBM Gaur et al., 2022a	Comparación entre modelos de caja negra e interpretables logrando una exactitud del 98.57 %; EBM mostró un rendimiento competitivo con transparencia estructural.
XGBoost + SHAP Lundberg et al., 2018	Modelo de ensamble basado en árboles explicado con valores SHAP para cuantificar contribuciones de características en tareas de predicción clínica, demostrando interpretabilidad consistente para factores de riesgo relacionados con diabetes.
AutoML + SHAP, LIME Hasan et al., 2024	Marco híbrido AutoML–XAI para predicción de riesgo de diabetes combinando modelos de ensamble con explicaciones SHAP y LIME, logrando una exactitud del 85 % mientras destaca glucosa e IMC como predictores dominantes.
Modelos ML + SHAP Khan et al., 2025	Estudio comparativo a gran escala evaluando múltiples modelos de aprendizaje automático para predicción de diabetes usando explicaciones SHAP para cuantificar la importancia de características en más de 70,000 muestras.
XGBoost + SHAP Huang et al., 2026	Modelo clínico basado en ultrasonido con explicación post-hoc vía SHAP, reportando un AUC superior a 0.90 para soporte a la decisión médica.
	Trabajos relacionados con DiabetIA
TabTransformer Flores-Custodio y Pancardo-García, 2025	Aplicación de arquitecturas basadas en atención en datos tabulares mexicanos, logrando una exactitud predictiva del 87.72 % con un enfoque orientado a la explicabilidad.
Transformer (Tabular) González y Nolasco, 2025	Arquitectura basada en atención aplicada para identificar nefropatía diabética en la población mexicana, representando un precedente directo utilizando el conjunto de datos DiabetIA.

Tabla 6.3. Comparación de métricas de rendimiento por clasificador.

Clasificador	***Accuracy***	***Precision***	***Recall***
Naïve Bayes Gaussiano	0.77	0.73	0.85
SVM	0.76	0.71	0.88
Árboles de decisión	0.72	0.73	0.69
Gradient boosting	0.78	0.74	0.86
Random forest	0.80	0.77	0.87
EBM	0.78	0.73	0.87

Tabla 6.4. Desglose de la matriz de confusión.

	Aciertos		**Errores**	
Clasificador	**TP**	**TN**	**FP**	**FN**
Naïve Bayes Gaussiano	186	149	32	70
SVM	191	140	27	79
Árboles de decisión	151	164	67	55
Gradient boosting	187	153	31	66
Random forest	195	145	28	69
EBM	189	150	29	69

Nota: TP = verdaderos positivos (casos detectados), TN = verdaderos negativos, FP = falsos positivos, FN = falsos negativos (casos omitidos).

Tabla 6.5. Correlación de Pearson con la variable objetivo (`e112`).

Variable clínica	**Coeficiente (r)**
Hipertensión esencial primaria (`essential_primary_hypertension`)	0.53
Edad al diagnóstico (`age_at_wx`)	0.46
Edad actual (`edad`)	0.42
Presión sistólica media (`fn_ta_systolic_mean`)	0.26
Presión diastólica media (`fn_ta_diastolic_mean`)	0.06
Obesidad (`obesity`)	0.10
Insuficiencia cardíaca (`heart_failure`)	0.04
Sexo masculino (`cs_sex_M`)	0.04
Trastornos de fluidos/electrolitos (`other_fluid_electrolyte_and_acid_base_disorders`)	0.04
Estatura media (`fn_height_mean`)	-0.02

Tabla 6.6. Comparativa interpretativa entre EBM y SHAP en la predicción de insuficiencia renal.

Caso	EBM	SHAP (Random forest)	Discusión
TP	Riesgo dominado por hipertensión y edad avanzada. Identifica factores clínicos clásicos.	Predicción impulsada por edad elevada, presión diastólica alta e hipertensión.	Coinciden en los factores clave; SHAP agrega el detalle de la presión diastólica.
TN	Acierto basado en presión arterial controlada y edad moderada (factores protectores).	Clasificación correcta sustentada por ausencia de hipertensión y presiones normales.	Coincidencia total en la identificación de factores protectores.
FP	Error motivado por presión sistólica alta, llevando a sobreestimar el riesgo.	Edad avanzada e hipertensión inducen una falsa predicción positiva (“falsa alarma”).	Coinciden en la sobreestimación de riesgo por perfil cardiovascular (presión/edad).
FN	Ausencia de hipertensión reduce indebidamente el puntaje, generando omisión del caso.	La falta de hipertensión actúa como una “barrera” que oculta otros síntomas.	Coincidencia en el subdiagnóstico por ausencia de hipertensión; es el sesgo más importante a corregir.
Global	Patrones estables de riesgo/protección; estructura aditiva clara.	Análisis granular e interactivo; revela contribuciones locales complejas.	Ambos modelos son coherentes, EBM ofrece estabilidad global y SHAP profundidad local.

Capítulo 7

Conclusiones y recomendaciones generales

7.1. Conclusiones

La presente investigación demuestra que la integración de la Inteligencia Artificial Explicable (XAI) permite el diagnóstico predictivo de insuficiencia renal en una herramienta de soporte clínico interpretable y auditable. A través del análisis comparativo entre el conjunto de datos del repositorio internacional UCI *CKD Dataset* y el conjunto de datos de México *DiabetIA*, fue posible evaluar tanto el rendimiento predictivo como la capacidad explicativa de distintos modelos de aprendizaje automático aplicados a la detección temprana de insuficiencia renal.

La metodología de esta tesis doctoral se desarrolló de manera estructurada a través de las siguientes publicaciones:

1. En el primer artículo (Capítulo 2) derivado de esta investigación se realizó un análisis univariado y descriptivo sobre el conjunto de datos CKD Dataset. Este estudio inicial permitió diagnosticar problemas críticos de la calidad de los datos, revelando que el $100\,\%$ de las variables presentaban valores faltantes, con inconsistencias particulares en variables clave como la diabetes mellitus (`dm`). Este hallazgo demostró la necesidad de implementar estrategias de imputación y normalización para mitigar sesgos y valores atípicos que habrían degradado la precisión y confiabilidad de cualquier modelo predictivo subsecuente.

2. El artículo posterior (Capítulo 3), con el fin de entender a fondo el funcionamiento interno del algoritmo Explainable Boosting Machine (EBM), se evaluó su comportamiento en el dataset de referencia de la *Cleveland Clinic Foundation* para enfermedades cardíacas. EBM demostró ser una herramienta eficaz alcanzando un desempeño equilibrado entre precisión (0.8710) y sensibilidad (0.8438). El análisis de sus casos de error (falsos negativos) evidenció que estos surgían por diferente aporte de las variables, lo que da sustento a esta investigación de que no basta solamente con predecir, sino que es necesario comprender las razones detrás de cada decisión algorítmica.

3. En el tercer artículo (Capítulo 4), una vez asimilado el funcionamiento de EBM, se aplicó sobre el *UCI CKD Dataset* de insuficiencia renal crónica, previamente analizado en el Capítulo 2. Se demostró que EBM el *trade-off* entre precisión y transparencia que fue uno de los objetivo iniciales de esta investigación. A nivel global, el algoritmo priorizó correctamente biomarcadores de alta relevancia médica como la gravedad específica urinaria (`sg`), la hemoglobina (`hemo`) y la creatinina sérica (`sc`), mientras que a nivel local proporciona una herramienta interpretativa capaz de desglosar detalladamente el riesgo individual de cada paciente.

4. Luego en el siguiente artículo (Capítulo 5) la investigación se aplicó al conjunto de datos mexicano DiabetIA. Se construyeron diversos modelos donde el algoritmo Random forest obtuvo el rendimiento predictivo óptimo. Para aminorar la opacidad de este modelo de caja negra, se integró el método de explicabilidad *post-hoc* LIME. Los resultados confirmaron que el modelo basaba sus decisiones en factores clínicamente coherentes (hipertensión y edad), pero la inspección granular de LIME permitió identificar fallos críticos en falsos negativos, evidenciando detalles clínicos de depender de métricas globales sin una auditoría local por paciente.

5. Finalmente (Capítulo 6), en el último artículo se conjuntó todo el conocimiento anterior sobre el dataset DiabetIA. Se implementó EBM y se añadió la explicabilidad a los modelos opacos usando SHAP. Esta integración demostró formalmente que la transparencia de EBM no compromete el rendimiento predictivo frente a algoritmos de caja negra. El análisis con SHAP funcionó como un mecanismo de auditoría clínica que descubrió una vulnerabilidad

en la ausencia de un diagnóstico explícito de hipertensión, ya que esto provocaba que los algoritmos subestimaran el riesgo real de daño renal, generando falsos negativos en pacientes con diabetes mellitus tipo 2.

De acuerdo con los resultados obtenidos de las cinco publicaciones, se presentan las siguientes conclusiones respecto a las hipótesis (Sección 1.6) planteadas en esta investigación:

- La hipótesis 1 ha sido aceptada, ya que los experimentos demostraron que el modelo intrínsecamente interpretable EBM superó el umbral métrico propuesto del $80\,\%$ de rendimiento, logrando exactitudes promedio y niveles de precisión superiores al $84\,\%$ en los diferentes escenarios de prueba. Asimismo, las pruebas estadísticas y de validación cruzada confirmaron un comportamiento de no-inferioridad (sin diferencias estadísticamente significativas) en comparación con los modelos obtenidos por Random forest y SVM, comprobando que los modelos transparentes pueden ser igual de precisos.
- La hipótesis 2: La hipótesis ha sido aceptada, ya que la incorporación de las técnicas*post-hoc* SHAP y LIME permitió cuantificar formalmente la contribución individual de las variables a nivel local y global. Mientras que los modelos predictivos convencionales aíslan el diagnóstico en una salida binaria, la aplicación de estas técnicas permitió auditar de forma estadísticamente significativa el peso específico de factores como la hipertensión arterial y la edad.

A pesar de las diferencias entre el CKD dataset (centrado en variables bioquímicas) y el conjunto de datos mexicano DiabetIA (influenciado por variables clínicas como la edad y la hipertensión), la metodología propuesta logró validar los resultados obtenidos con 2 cohortes diferentes. El uso de un conjunto de datos representativo de la población mexicana aporta relevancia al proyecto, ya que los modelos creados pueden ser utilizados por personal médico como apoyo a la detección temprana de insuficiencia renal.

7.2. Recomendaciones

A partir de los resultados obtenidos, se plantean las siguientes líneas de investigación futuras:

- Debido a que la hipertensión es la variable más influyente en predicción de insuficiencia renal, se propone optimizar la sensibilidad del modelo mediante ingeniería de características para detectar daño renal en pacientes normotensos.
- Dado que el dataset DiabetIA representa una población del estado de Michoacán, es imperativo recolectar y analizar un conjunto de datos propio del estado de Tabasco. Esto permitirá capturar las particularidades genéticas, dietéticas y sociodemográficas de la población del sureste mexicano, mejorando la precisión local de los modelos.
- Implementar pruebas entre diferentes instituciones de salud de la región para garantizar que las explicaciones de SHAP y EBM sean generalizables y no dependan de un único centro hospitalario.
- Desarrollar una herramienta visual que permita a médicos y especialistas interactuar con las explicaciones del modelo en tiempo real durante la consulta.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Diseño de modelos explicativos de aprendizaje automático para detención de insuficiencia renal
Autor:	Simón Javier Hernández Gaspar
ORCID:	https://orcid.org/0009-0003-1095-713X0009-0003-1095-713X
Resumen:	La presente tesis aborda el desafío de la detección temprana de complicaciones renales en pacientes con diabetes tipo 2 mediante el uso de Inteligencia Artificial Explicable (XAI). El objetivo central fue evaluar y comparar la eficacia de modelos intrínsecamente interpretables frente a explicaciones post-hoc para garantizar diagnósticos precisos y transparentes en el entorno clínico. La investigación se desarrolló en tres fases críticas. En la primera, se validó el algoritmo Explainable Boosting Machine (EBM) utilizando un conjunto de datos internacional (UCI) de enfermedad renal crónica, logrando una precisión del 99 % e identificando marcadores bioquímicos clave como la gravedad específica y la creatinina sérica. En la segunda fase, se analizó el dataset DiabetIA (42,182 instancias de población mexicana), donde el modelo Random Forest alcanzó un rendimiento del 80 %. Mediante la técnica LIME, se demostró que en el contexto nacional, la hipertensión esencial y la edad son los factores de riesgo dominantes. Finalmente, se realizó una auditoría de interpretabilidad cruzada comparando EBM y SHAP. Los resultados revelaron una vulnerabilidad crítica en los modelos: un sesgo que tiende a subestimar el riesgo en pacientes normotensos (Falsos Negativos), donde la ausencia de hipertensión oculta otras señales de alarma. Se concluye que la integración de XAI no solo iguala el rendimiento de los algoritmos de çaja negra", sino que actúa como una herramienta de auditoría clínica indispensable para reducir omisiones diagnósticas y fortalecer la medicina de precisión en el sistema de salud mexicano.
Palabras clave:	Inteligencia Artificial Explicable, EBM, SHAP, LIME, Nefropatía Diabética, Auditoría Clínica.
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology*, *20*(1), 105. https://doi.org/10.1186/s13000-025-01686-3

Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, *24*, 100631. https://doi.org/10.1016/j.imu.2021.100631

Arias Rodríguez, M., Aljama García, P., Egido de los Ríos, J., Lamas Peláez, S., Praga Terente, M., & Serón, D. (2013). *Nefrología clínica*. Editorial Médica Panamericana.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters*, *35*(2), 131-150. https://doi.org/10.1007/s11063-011-9208-y

Avci, E., Karakus, S., Ozmen, O., & Avci, D. (2018). Performance comparison of some classifiers on Chronic Kidney Disease data. *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, 1-4. https://doi.org/10.1109/ISDFS.2018.8355392

Bagnato, J. I. (2020). Aprende Machine Learning en Español: Teoría + Práctica Python. https://www.aprendemachinelearning.com

Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific Reports*, *12*, 8377. https://doi.org/10.1038/s41598-022-12316-z

Barakat, N., & Diederich, J. (2008). Eclectic Rule-Extraction from Support Vector Machines. *International Journal of Computer and Information Engineering*, *2*(5), 1672-1675.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bisong, E. (2019). The Multilayer Perceptron (MLP). En *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 401-405). Apress. https://doi.org/10.1007/978-1-4842-4470-8_31

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth; Brooks/Cole.

Carneiro, T., Medeiros, R. M., Nepomuceno, D. B., Bian, G. B., da Costa, V. H., & de Albuquerque, P. S. (2018). Performance Analysis of Google Colaboratory as a Tool for Machine Learning. *Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1234-1239. https://doi.org/10.1109/CSCI46756.2018.00235

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2788613

Cohen, L., Mansour, Y., Moran, S., & Shao, H. (2024). Probably Approximately Precision and Recall Learning [Disponible en: https://arxiv.org/abs/2411.13029]. *arXiv preprint arXiv:2411.13029*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, *9*, 109. https://doi.org/10.1186/s40537-022-00657-5

Deitel, P., & Deitel, H. (2019). *Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and the Cloud* (1.ª ed.). Pearson.

Dery, L. (2026). Heart disease dataset (heart.csv) [Dataset disponible en el repositorio EDA-course]. https://github.com/nlihin/EDA-course/blob/main/datasets/heart.csv

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9

Dirección General de Comunicación Social, UNAM. (2020, septiembre). Enfermedades del corazón, pandemia permanente.

Dirección General de Epidemiología. (2025). *Informe trimestral del Sistema de Vigilancia Epidemiológica Hospitalaria de Diabetes Mellitus Tipo 2 (SVEHDMT2), tercer trimestre 2025* (inf. téc.). Secretaría de Salud, Gobierno de México. https://www.gob.mx/salud/documentos/diabetes-mellitus-tipo-2-hospitalaria-2025

Duval, A. (2019). *Explainable Artificial Intelligence (XAI)* (Scholarly Report). Mathematics Institute, The University of Warwick.

El Heraldo de México. (2022). Supera Tabasco en 4 años más de mil muertes renales. https://oem.com.mx/elheraldodetabasco/ciencia-y-salud/supera-tabasco-en-4-anos-mas-de-mil-muertes-renales-19406978

Erbani, J., Portier, P.-É., Egyed-Zsigmond, E., & Nurbakova, D. (2024). Confusion Matrices: A Unified Theory. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2024.10769075

Federación Mundial del Corazón. (2026). ¿Qué son las enfermedades cardiovasculares? [Accedido: 08-abr-2026].

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer. https://doi.org/10.1007/978-3-319-98074-4

Flores-Custodio, O., & Pancardo-García, P. (2025). A Transformer Network Model for the Diagnosis of Chronic Ischemic Heart Disease in Patients with Type 2 Diabetes Mellitus. *Artificial Intelligence – COMIA 2025*, *2554*, 245-260. https://doi.org/10.1007/978-3-031-97913-2_19

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Ganie, S. M., Pramanik, P. K. D., & Zhao, Z. (2025). Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Scientific Reports*, *15*(13912). https://doi.org/10.1038/s41598-025-97547-6

Gaur, L., Biswas, M., Bakshi, S., Gupta, P., Si, T., Mallik, S., & Maulik, U. (2022a). An Integrated Model to Evaluate the Transparency in Predicting Chronic Kidney Disease Using a Trio-Embedded Explainable Model [SSRN Preprint]. https://doi.org/10.2139/ssrn.4129888

Gaur, L., Biswas, M., Bakshi, S., Gupta, P., Si, T., Mallik, S., & Maulik, U. (2022b). An Integrated Model to Evaluate the Transparency in Predicting Chronic Kidney Disease Using a Trio-Embedded Explainable Model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4129888

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2.ª ed.). O'Reilly Media.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3.ª ed.). O'Reilly Media.

Ghosh, P., Shamrat, F. M. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020). Optimization of prediction method of chronic kidney disease using machine learning algorithm. *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1-6. https://doi.org/10.1109/iSAI-NLP51646.2020.9376787

Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, *14*(3687). https://doi.org/10.1038/s41598-024-54375-4

Gogoi, P., & Valan, J. A. (2025). Interpretable Machine Learning for Chronic Kidney Disease Prediction: A SHAP and Genetic Algorithm-Based Approach. *Biomedical Materials & Devices*, *3*, 1384-1402. https://doi.org/10.1007/s44174-024-00262-5

Gokiladevi, M., Santhoshkumar, S., & Varadarajan, V. (2022). Machine learning algorithm selection for chronic kidney disease diagnosis and classification. *Malaysian Journal of Computer Science*, 102-115. https://doi.org/10.22452/mjcs.sp2022no1.8

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2014). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1). https://doi.org/10.1080/10618600.2014.907095

Gómez García, A., López Pineda, A., García Velázquez, L. M., Flores Garrido, M., Álvarez Aguilar, C., & Figueroa Mora, K. M. (2023). Base de Datos DiabetIA [Conjunto de datos]. https://repositorio-salud.conacyt.mx/jspui/handle/1000/296

González, L. R. T. M., & Nolasco, J. A. H. (2025). Modelo Transformer para la identificación de la nefropatía diabética en pacientes Mexicanos. *Research in Computing Science*, *154*(5), 87-101. https://www.rcs.cic.ipn.mx/2025_154_5/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gunning, D. (2017). *Explainable artificial intelligence (xAI)* (inf. téc.). Defense Advanced Research Projects Agency (DARPA).

Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)* (inf. téc.). Defense Advanced Research Projects Agency (DARPA).

Hara, S., & Hayashi, K. (2016). Making Tree Ensembles Interpretable.

Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. (2024). Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information*, *16*(1), 7. https://doi.org/10.3390/info16010007

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second). Springer Science & Business Media.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hernández-Gaspar, S. J., Hernández-Ocaña, B., Hernández-Torruco, J., & Chávez-Bosquez, O. (2025). Inteligencia artificial explicable para el análisis de pacientes con insuficiencia renal crónica. *Nova Scientia*, *17*, 1-22. https://doi.org/10.21640/ns.v17iSpecialIa.3630

Huang, Y., Su, S., Chen, J., Zhang, X., Tan, K., & Guo, Q. (2026). An ultrasound-based machine learning model for predicting pelvic adhesions: A SHAP-enhanced XGBoost approach. *DIGITAL HEALTH*, *12*. https://doi.org/10.1177/20552076261416797

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

IDF. (2021). *International Diabetes Federation Diabetes Atlas* (10th). International Diabetes Federation. https://diabetesatlas.org/

Instituto Mexicano del Seguro Social. (2013). *Acuerdo de cooperación técnica IMSS- OPS/OMS* [Disponible en www.imss.gob.mx/NR/rdonlyres/…/convenio_cooperacion.pdf (fecha de consulta 03/06/13)].

Instituto Mexicano del Seguro Social. (2025). DiabetIA: Conjunto de datos generado a partir del Sistema de Informática de Medicina Familiar (SIMF) en Michoacán [Información derivada de expedientes clínicos de pacientes con diabetes mellitus tipo 2, siguiendo criterios de la Clasificación Internacional de Enfermedades (CIE)].

Instituto Nacional de Salud Pública. (2020). Enfermedad renal crónica en México. https://www.insp.mx/avisos/5296-enfermedad-renal-cronica-mexico.html

Instituto Nacional de Salud Pública. (2021). *Encuesta Nacional de Salud y Nutrición 2021 sobre Covid-19. Resultados Nacionales* (inf. téc.). INSP. https://ensanut.insp.mx/

International Diabetes Federation. (2025). *IDF Diabetes Atlas — Datos y cifras sobre la diabetes* [Accedido: 19-nov-2025]. International Diabetes Federation. https://idf.org/es/about-diabetes/diabetes-facts-figures/

Itoh, T., Kumar, A., Klein, K., & Kim, J. (2017). High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *Journal of Visual Languages & Computing*, *43*, 1-13. https://doi.org/10.1016/j.jvlc.2017.03.001

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2021). Leveraging SHAP and LIME for Enhanced Explainability in AI-Driven Diagnostic Systems. *International Journal of AI and ML*, *2*(3). https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/81

Kaur, C., Kumar, M. S., Anjum, A., Binda, M. B., Mallu, M. R., & Al Ansari, M. S. (2023). Chronic Kidney Disease Prediction Using Machine Learning. *Journal of Advances in Information Technology*, *14*(2), 384-391. https://www.jait.us/uploadfile/2023/JAIT-V14N2-384.pdf

KDIGO. (2022). *Kidney Disease Improving Global Outcomes - Clinical practice guideline for diabetes management in chronic kidney disease* (inf. téc.). https://kdigo.org/wp-content/uploads/2022/10/KDIGO-2022-Clinical-Practice-Guideline-for-Diabetes-Management-in-CKD.pdf

Khan, M., Ahmed, S., & Rahman, M. (2025). A comparative study of explainable machine learning models with Shapley values for diabetes prediction. *Healthcare Analytics*, *7*, 100390. https://doi.org/10.1016/j.health.2025.100390

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*. https://doi.org/10.48550/arXiv.1707.01154

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350-1371. https://doi.org/10.1214/15-AOAS848

Levey, A. S., Atkins, R., Coresh, J., Cohen, E. P., Collins, A. J., Eckardt, K. U., Nahas, M. E., Jaber, B. L., Jadoul, M., Levin, A., Powe, N. R., Rossert, J., Wheeler, D. C., Lameire, N., & Eknoyan, G. (2007). Chronic kidney disease as a global public health problem: Approaches

and initiatives – a position statement from Kidney Disease Improving Global Outcomes. *Kidney International*, *72*(3), 247-259. https://doi.org/10.1038/sj.ki.5002343

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and Understanding Neural Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 188-197.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3.ª ed., Vol. 793). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Lundberg, S. M., Erion, G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 4768-4777.

Lundberg, S. M., & Lee, S.-I. (2017b). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 4765-4774. https://arxiv.org/abs/1705.07874

Maron, M. E. (1961). Automatic indexing: An experiment with a statistical model. *Journal of the ACM*, *8*(3), 404-417. https://doi.org/10.1145/321075.321084

McKinney, W. (2010). Data Structures for Statistical Computing in Python. En S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2014). *Introducción a la Probabilidad y Estadística* (14.ª ed.). CENGAGE Learning.

Mengle, S. S., & Gurmendez, M. (2019). *Mastering Machine Learning on AWS*. Packt Publishing.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269*. https://arxiv.org/abs/1706.07269

Mishra, A. (2021). *Mastering Machine Learning on AWS*. Packt Publishing.

Mitchell, T. M. (2019). *Machine Learning* (2.ª ed.). McGraw-Hill.

Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book

Myers, R. H., Walpole, R. E., & Myers, S. L. (2012). *Probabilidad y Estadística para ingeniería y ciencias* (9.ª ed.). PEARSON EDUCACIÓN.

Nogueira, A. R., Ferreira, C. A., & Gama, J. (2018). Improving acute kidney injury detection with conditional probabilities. *Intelligent Data Analysis*, *22*(6), 1355-1374. https://doi.org/10.3233/IDA-173626

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability [arXiv preprint arXiv:1909.09223]. https://doi.org/10.48550/arXiv.1909.09223

Organización Mundial de la Salud. (2021, junio). Enfermedades cardiovasculares.

Organización Panamericana de la Salud. (2021). Carga de enfermedades renales. https://www.paho.org/es/enlace/carga-enfermedes-renales

Organización Panamericana de la Salud. (2026). *ENLACE: Portal de Datos sobre Enfermedades No Transmisibles, Salud Mental, y Causas Externas* [Accedido en 2026]. https://www.paho.org/es/enlace/carga-enfermedes-renales

Organización para la Cooperación y el Desarrollo Económicos. (s.f.). Acerca de la OCDE [Accedido: 08-abr-2026].

Pal, S. (2023). Chronic Kidney Disease Prediction Using Machine Learning Techniques. *Biomedical Materials & Devices*, *1*, 534-540. https://doi.org/10.1007/s44174-022-00027-y

Paul, J. (2024). LIME and SHAP interpretability for medical AI systems. https://doi.org/10.13140/RG.2.2.14652.45443

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Poveda García, D. A., Ramírez Guzmán, M. E., Fernández Ordóñez, Y. M., & García. López, E. F. (2024). Implementación del método de optimización de máquinas de soporte vectorial en R. *Tecnura*, *28*(80), 99-118. https://doi.org/10.14483/22487638.20326

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2020). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991-21002. https://doi.org/10.1109/ACCESS.2019.2963053

Rahimiaghdam, S., & Alemdar, H. (2025). MindfulLIME: A Stable Solution for Explanations of Machine Learning Models with Enhanced Localization Precision. *arXiv preprint arXiv:2503.20758*. https://arxiv.org/abs/2503.20758

Raihan, M. J., Khan, M. A., Kee, S. H., & Nahid, A. A. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, *13*(1), 6263. https://doi.org/10.1038/s41598-023-33525-0

Rajani, N. F., & Mooney, R. J. (2018). Ensembling Visual Explanations. En *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 155-172). Springer. https://doi.org/10.1007/978-3-319-98131-4_7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You¿': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rosenbaum, L., Hinselmann, G., & Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. *Journal of Cheminformatics*, *3*, 11. https://doi.org/10.1186/1758-2946-3-11

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Dataset. https://doi.org/10.24432/C5G020

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Learning*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Russell, S., & Norvig, P. (2022). *Inteligencia artificial: Un enfoque moderno* (4.ª ed.). Pearson Educación.

Salman, M., Al-Saedi, A., & Al-Saedi, H. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2*(1), 1-10. https://doi.org/10.5281/zenodo.10812345

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Secretaría de Salud. (2025). *Informe Trimestral de Vigilancia Epidemiológica Hospitalaria de Diabetes Mellitus Tipo 2 (3er Trimestre 2025)* (Informe Trimestral) (Accedido: 19-nov-2025). Gobierno de México. https://www.gob.mx/cms/uploads/attachment/file/1029886/InformeTrimestralSVEDMT2-3erTrim-2025.pdf

Shaikh, A. S., Samant, R. M., Patil, K. S., Patil, N. R., & Mirkale, A. R. (2023). Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain. *International Journal of Computer Applications*, *185*(45), 18-23. https://www.ijcaonline.org/archives/volume185/number45/32992-2023923263/

Shankar, S., Verma, S., Elavarthy, S., Kiran, T., & Ghuli, P. (2020). Analysis and prediction of chronic kidney disease. *International Research Journal of Engineering and Technology*, *7*(5), 2395-0056. https://www.irjet.net/archives/V7/i5/IRJET-V7I5868.pdf

Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. *Diagnostics*, *12*(1), 116. https://doi.org/10.3390/diagnostics12010116

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second). MIT Press.

Swapna, G., Acharya, U. R., VinithaSree, S., & Suri, J. S. (2013). Automated detection of diabetes using higher-order spectral features extracted from heart rate signals. *Intelligent Data Analysis*, *17*(2), 309-326. https://doi.org/10.3233/IDA-130580

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd). Graphics Press.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

World Kidney Day. (2022). Campaña de prevención renal 2022. https://www.worldkidneyday.org/events/campana-de-prevencion-renal-2022/

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond Sparsity: Tree Regularization of Deep Models for Interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 1670-1678.

Wu, Y., Zhang, L., Bhatti, U. A., & Huang, M. (2023). Interpretable Machine Learning for Personalized Medical Recommendations: A LIME-Based Approach. *Diagnostics*, *13*(16), 2681. https://doi.org/10.3390/diagnostics13162681

Yagin, F. H., Aygun, U., Colak, C., Alkhalifa, A. K., Alzakari, S. A., & Aghaei, M. (2025). Accuracy is not enough: Explainable boosting machine model and identification of candidate biomarkers for real-time sepsis risk assessment in the emergency department. *BMC Emergency Medicine*, *25*(1), 248. https://doi.org/10.1186/s1471-227X-25-248

Yousefi, L., & Tucker, A. (2022). Identifying latent variables in dynamic Bayesian networks with bootstrapping applied to prediction of type 2 diabetes complications. *Intelligent Data Analysis*, *26*(2), 501-524. https://doi.org/10.3233/IDA-205570

Zhou, Y., & Hooker, G. (2016). Interpreting Models via Single Tree Approximation.