ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

EVALUACIÓN Y APLICACIÓN DE ALGORITMOS DE INTELIGENCIA ARTIFICIAL EXPLICADA PARA APOYAR LA TOMA DE DECISIONES MÉDICAS EN LA SALUD FETAL

EVALUACIÓN Y APLICACIÓN DEL ALGORITMO LIME DE INTELIGENCIA ARTIFICIAL EXPLICADA PARA APOYAR LA TOMA DE DECISIONES MÉDICAS EN LA SALUD FETAL

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SOFTWARE

ISMAEL SEBASTIÁN RIVAS HIDALGO

DIRECTOR: EDISON FERNANDO LOZA AGUIRRE, PhD.

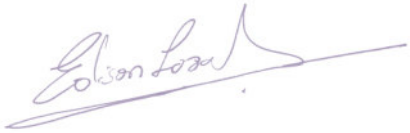
DMQ, marzo de 2022

CERTIFICACIONES

Yo, ISMAEL SEBASTIÁN RIVAS HIDALGO declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

ISMAEL SEBASTIÁN RIVAS HIDALGO

Certifico que el presente trabajo de integración curricular fue desarrollado por ISMAEL SEBASTIÁN RIVAS HIDALGO, bajo mi supervisión.



EDISON FERNANDO LOZA AGUIRRE
DIRECTOR

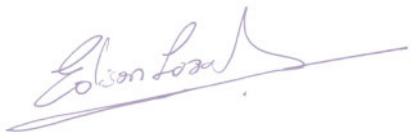
Certificamos que revisamos el presente trabajo de integración curricular.

NOMBRE_REVISOR1	**NOMBRE_REVISOR2**
REVISOR1 DEL TRABAJO DE INTEGRACIÓN CURRICULAR	**REVISOR2 DEL TRABAJO DE INTEGRACIÓN CURRICULAR**

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

ISMAEL SEBASTIÁN RIVAS HIDALGO



EDISON FERNANDO LOZA AGUIRRE

DEDICATORIA

A mis padres Homero y Hermita, quienes han sido el pilar fundamental de mi desarrollo como ser humano; todos mis logros se los debo a ustedes.

A mis hermanos, por preocuparse por mí en todo momento y estar siempre presentes cuando lo necesito.

A mis abuelos, por su cariño y afecto; en especial a mi abuelo Germán, por inculcarme el sentido de responsabilidad y la importancia del esfuerzo constante.

A mi tía Eulalia, a quien considero como una segunda madre.

Y a todas las personas que se esmeran por alcanzar sus objetivos, este es un testimonio de que todo esfuerzo rinde frutos.

AGRADECIMIENTO

Agradezco a mi familia por todo su sacrificio para el logro de mis metas.

A la Escuela Politécnica Nacional por la invaluable formación académica y personal.

Al doctor Edison Loza, quien más que un tutor, ha sido un amigo.

A Absalón y su familia, por su apoyo incondicional y desinteresado.

A mis amigos, por llenar de aventuras este viaje.

Gracias a todos.

ÍNDICE DE CONTENIDO

RESUMEN

Este trabajo de titulación responde a la necesidad de entendimiento en las decisiones de los modelos de AI en la clasificación de la salud fetal, para esto se ha propuesto una comparación de algoritmos de aprendizaje automático y la implementación de un modelo demostrador de Inteligencia Artificial Explicada a través de LIME que permita apoyar la toma de decisiones sobre diagnósticos médicos. El proceso de elaboración del proyecto ha sido apoyado por la metodología CRISP-DM, aprovechando su enfoque orientado al análisis de datos e implementación de modelos inteligentes. Los resultados de su ejecución indican la falta de literatura respecto a los modelos de XAI en obstetricia; además, luego de la comparación y evaluación contextualizada de los modelos de AI: SVM, ANN y Random Forest, es este último el que presenta los mejores resultados. Respecto al algoritmo de XAI, las características influyentes en las decisiones determinan que para una instancia clasificada como “Normal”, valores bajos en aceleraciones y desaceleraciones prolongadas apoyan esta decisión. Para una instancia clasificada como “Sospechosa”, el porcentaje de tiempo prolongado con variabilidad anormal a largo plazo soporta esta clasificación. Finalmente, en una instancia clasificada como “Patológica”, las desaceleraciones prolongadas y severas en valores mínimos contradicen este resultado, por lo tanto, valores elevados podrían advertir una patología en la salud del feto. Estas explicaciones han sido validadas por un experto en el campo obstétrico a través del Modelo de Aceptación de Tecnología (TAM) en lo que respecta a la facilidad de uso y la utilidad percibida.

PALABRAS CLAVE: XAI, toma de decisiones, LIME, salud fetal, explicación, interpretación.

ABSTRACT

This work answers the need for understand the decisions of AI models in the classification of fetal health. In order to do this, a comparison of machine learning algorithms and the implementation of a demonstrator model of Explained Artificial Intelligence through LIME has been proposed to support decision making on medical diagnoses. The project elaboration process has been supported by the CRISP-DM methodology, taking advantage of its approach oriented to data analysis and intelligent models' implementation. The results of its execution indicate the lack of literature regarding XAI models in obstetrics; moreover, after the comparison and contextualized evaluation of the AI models: SVM, ANN and Random Forest, it is the latter that grants the best results. Regarding the XAI algorithm, the decision-influencing characteristics determine that for an instance classified as "Normal", low values in accelerations and prolonged decelerations support this decision. For an instance classified as "Suspicious", the percentage of prolonged time with abnormal long-term variability supports this classification. Finally, for an instance classified as "Pathological", prolonged and severe decelerations at minimum values contradict this result, therefore, higher values could warn of a pathology in the fetal health. These explanations have been validated by an expert in the obstetric field through the Technology Acceptance Model (TAM) in terms of ease of use and perceived usefulness.

KEYWORDS: XAI, decision making, LIME, fetal health, explanation, interpretation.

1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

Los problemas de salud en el ser humano se pueden presentar desde antes de su nacimiento, pues todos los embarazos presentan un riesgo en menor o mayor medida. El diagnóstico de estos problemas es dependiente del criterio médico con el que cada caso sea analizado, lo que afecta directamente la salud del feto. Tomar una decisión sobre el tratamiento que se debe aplicar involucra una situación comprometedora, sobre todo si se consideran las consecuencias de un error en el análisis del caso.

En las últimas décadas, la medicina ha encontrado en la inteligencia artificial (AI, por sus siglas en inglés) un propulsor en la búsqueda de soluciones. Así, mediante algoritmos de aprendizaje automático, tareas tan específicas como el diagnóstico de pacientes se ven favorecidas por su capacidad de predicción basada en los datos. La AI ayuda a reducir el riesgo de cometer errores de diagnóstico, al minimizar la intervención humana en la evaluación de la salud del paciente.

A pesar de los beneficios que otorga la AI en el campo de la medicina, su funcionamiento aún presenta incertidumbre para los médicos, limitando su adopción. La AI se presenta entonces como una “caja negra” que esconde el proceso de aprendizaje para la generación de resultados.

Es así, que la inteligencia artificial explicada (XAI, por sus siglas en inglés) busca solventar esta problemática, exponiendo las decisiones detrás del algoritmo. Si se conocen las variables que influyen en el resultado, es posible determinar los puntos críticos que permiten formular una solución más efectiva. Así, esta información puede ser aprovechada por los médicos para incrementar el número de decisiones correctas en la elaboración de tratamientos médicos.

Hoy por hoy, la AI permite determinar la salud de un feto mediante algoritmos de predicción. Sin embargo, no existe la posibilidad de conocer los factores representativos del modelo que cumple este objetivo. La XAI, por su parte, permitiría determinar la influencia de factores como la frecuencia cardíaca fetal o las contracciones uterinas en la salud del feto, reduciendo la incertidumbre en la elaboración de diagnósticos médicos.

Dadas las consideraciones anteriores, se propone implementar un modelo demostrador de XAI, a través de LIME, que permita apoyar la toma de decisiones sobre diagnósticos médicos.

1.1 Objetivo general

Evaluar y aplicar el algoritmo de Inteligencia Artificial Explicada LIME para el apoyo en la toma decisiones médicas para la salud fetal.

1.2 Objetivos específicos

1. Realizar una Revisión Sistemática de Literatura referente a la XAI y sus implicaciones en la medicina.
2. Efectuar una evaluación comparativa entre algoritmos de aprendizaje automático para la predicción de la salud fetal.
3. Desarrollar un modelo de XAI de tipo LIME para soportar la toma de decisiones médicas en la salud fetal.
4. Validar la capacidad explicativa de los modelos con profesionales de la salud especializados en el área de la obstetricia.

1.3 Alcance

En primera instancia, se determinarán los objetivos de negocio identificando las necesidades de los profesionales de la salud para mejorar el diagnóstico respecto a la salud fetal. Se realizará una revisión sistemática de literatura que permita conocer los estudios de XAI en el campo de la medicina.

A continuación, se obtendrán los datos que servirán de entrada para los algoritmos de AI que se considerarán en esta investigación. Estos datos serán examinados minuciosamente con el fin de obtener información significativa acerca del problema. Para hacer uso de los datos obtenidos, estos pasarán por un proceso de limpieza y formateo que permita identificar datos erróneos, faltantes o incompletos que alteren la precisión del modelo a implementar.

Se llevará a cabo un proceso de selección para determinar los algoritmos de aprendizaje automático que permitirán clasificar la salud fetal con los datos de entrada. Posteriormente, se diseñará el plan de prueba para identificar la precisión de los algoritmos de aprendizaje automático mediante métricas de rendimiento. Adicionalmente, se considerará una separación de datos en conjuntos de entrenamiento y prueba.

Los modelos serán construidos con los parámetros requeridos por cada algoritmo y sus salidas serán evaluadas según el plan de prueba definido. Una vez obtenidos los resultados de los modelos AI, se construirán el modelo de XAI con LIME. La evaluación de los

resultados tiene la finalidad de seleccionar el algoritmo que cumpla con los objetivos del negocio y aporte valor en la toma de decisiones. Esta evaluación se llevará a cabo con la ayuda de un profesional de la salud especializado en el área de obstetricia.

Finalmente, se diseñará un plan de despliegue para los modelos de los algoritmos de aprendizaje automático y XAI. Para esto, se indicarán estrategias de implementación que permitan presentar los resultados a los interesados de manera legible como un sistema de información integral.

1.4 Marco teórico

1.4.1 Inteligencia Artificial Explicable

Hoy por hoy, la AI representa una de las herramientas más útiles e innovadoras en la toma de decisiones en ambientes de alta complejidad [1]. Esto se debe principalmente a su naturaleza basada en los datos y su capacidad para derivar un resultado de valor. Este resultado, como producto final de su elaboración minuciosa, parte desde predicciones simples hasta clasificadores de múltiples etiquetas, pues vasto es el estudio de la AI y la implementación de sus modelos [2]. Sin embargo, en la mayoría de los casos no existe la posibilidad de conocer aquellos factores representativos de aquel modelo que cumple el objetivo final para el que fue diseñado, pero que deja interrogantes en el camino.

Con esta premisa en mente, se idea la Inteligencia Artificial Explicada (XAI, por sus siglas en inglés) que como su nombre lo indica, tiene por objetivo explicar aquellos factores que parecen estar ocultos a simple vista en los modelos convencionales y afectan en mayor o menor medida el resultado final [3]. Esto se traduce en sistemas que podrían aprovechar el proceso de diseño e implementación de los modelos de AI para extraer los factores influyentes en la problemática a la que buscan responder, aumentando considerablemente la importancia de su aplicación para la toma de decisiones [4].

La relación entre la AI convencional y la explicada se puede distinguir en los enfoques propios de la implementación de la XAI en función de la naturaleza de los algoritmos de aprendizaje automático. Los principales enfoques son dos:

- Modelos intrínsicamente explicables que son diseñados con la intención de transparentar el proceso de aprendizaje y predicción, en donde no es necesaria una implementación adicional para cumplir este objetivo [5]. La Figura 1 representa la estructura de este tipo de modelos.

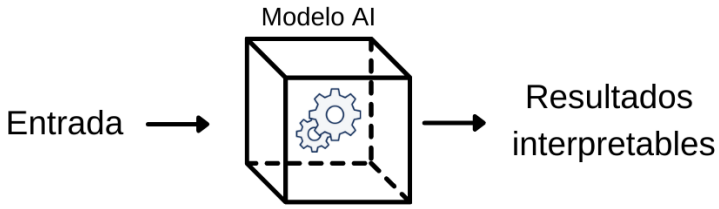


Figura 1. Estructura de un modelo intrínsecamente explicable.

- Modelos cuya naturaleza se basa en el resultado final, en donde los algoritmos de aprendizaje automático no abordan explicaciones de ningún tipo durante el proceso de entrenamiento. Estos algoritmos representan el término “black box” por la falta de transparencia en el proceso de modelización, generalmente tienen por objetivo la precisión y velocidad de ejecución [6]. En este caso es necesario un componente interpretativo adicional y, muchas veces, independiente; suele tomar el nombre de modelo “post-hoc” [7]. La estructura de este modelo se ilustra en la Figura 2.

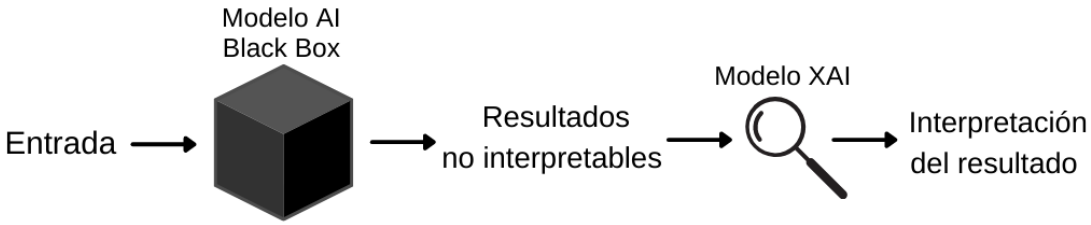


Figura 2. Estructura de un modelo post-hoc.

1.4.2 Principios de la Inteligencia Artificial Explicada

Independientemente de los enfoques en los que la XAI sea implementada, los principios que constituyen su naturaleza no varían. Con la finalidad de reducir la incertidumbre en la concepción de XAI, el US National Institute of Standards and Technology (NIST) ha desarrollado una serie de principios que deben cumplirse para que una tecnología sea considerada como un sistema de XAI [8]. Estos principios describen las propiedades fundamentales de la XAI y reducen la brecha en el entendimiento provocada por la naturaleza multidisciplinar de sus aplicaciones. Cada principio se describe a continuación:

Explicación

Es el principio que da nombre a esta técnica e indica la obligatoriedad de un algoritmo de AI a dar una explicación fundamentada del producto de salida. Es decir, el modelo de aprendizaje automático debe estar apoyado en evidencia para justificar su funcionamiento;

sin embargo, no se asume que dicha evidencia sea verídica, precisa o interpretable, pues cada principio de XAI es independiente y no influye en los demás.

Para ahondar en este principio, es posible describir los tipos de explicaciones que tienen por finalidad agregar interpretabilidad a un modelo de AI. Si bien, parecería correcto pensar que las explicaciones se categorizan únicamente a partir de los algoritmos que se emplean o los datos que se analizan, la realidad es que las explicaciones también se manejan acorde con los requerimientos de la problemática, donde incluso caben las necesidades del usuario y el nivel explicativo que estas requieran [9].

Significado

Este principio hace énfasis en el usuario al que se dirige el sistema de AI desarrollado, el valor que aporta la explicación del algoritmo está directamente relacionado con la capacidad del receptor para interpretarlo y entenderlo [10]. Sin embargo, es evidente que no todos los usuarios interpretan los resultados de la misma manera, pues la experiencia y el conocimiento previo pueden influir en su capacidad de abstraer la información [11]. Este principio remite la necesidad de producir explicaciones considerando a la mayor cantidad de usuarios posible.

Exactitud

Aquella evidencia que soporta la explicación dada por el sistema de XAI no ha sido probada para comprobar su fidelidad, es aquí donde se hace presente este principio. El estudio de NIST resalta la diferencia entre precisión en la explicación y precisión en la decisión [8], siendo esta última tradicionalmente utilizada con los sistemas de “caja negra” para evaluar la capacidad de predicción o clasificación de un algoritmo de AI [12]. Por su parte, la precisión en la explicación de los sistemas de XAI indica el nivel de interpretabilidad que ofrece dicho modelo sin dejar de lado la subjetividad del receptor que lo utiliza. Este aspecto dificulta la implementación de métricas de exactitud en la explicación de modelos y limita su uso estandarizado para establecer comparaciones, al contrario de lo que sucede con las métricas de exactitud en modelos no explicativos [13].

Límites del conocimiento

El último de los principios de los sistemas de XAI se refiere a la necesidad de que un sistema de IA identifique y declare los límites de conocimiento a los que está sujeto, con la finalidad de mantener la confianza en sus decisiones y evitar cualquier disconformidad por la existencia de resultados injustificados. De la misma forma en la que un algoritmo aprende en función de los datos determinados durante la fase de entrenamiento, el modelo

generado espera una entrada coherente con lo aprendido [14]. Sin embargo, en el caso de que el sistema desborde sus límites de conocimiento, este debe ser capaz de reconocer esta situación y otorgar decisiones congruentes con lo acontecido.

Estos principios permiten delimitar la concepción de Inteligencia Artificial Explicada y diseminarla en un ámbito multidisciplinar con mayor facilidad.

1.4.3 Métodos y algoritmos de XAI

Los algoritmos utilizados para implementar la XAI se pueden categorizar en dos grandes grupos según su enfoque respecto al alcance de la explicación:

Interpretación global del modelo

Una interpretación global permite entender el comportamiento de todo el modelo en general. Los patrones y reglas que se extrajeron durante el entrenamiento del modelo responden directamente a sus decisiones y estas impactan de la misma forma en todas las muestras de datos [15]. Este nivel de interpretabilidad profundiza en los elementos aprendidos, como los pesos en la definición de un algoritmo, para tratar de determinar la efectividad de un modelo o encontrar deficiencias que requieran alguna corrección [16]. Sin embargo, no busca indicarle al usuario cuáles son las características importantes o insignificantes que explican las decisiones tomadas por el modelo.

Interpretación local del modelo

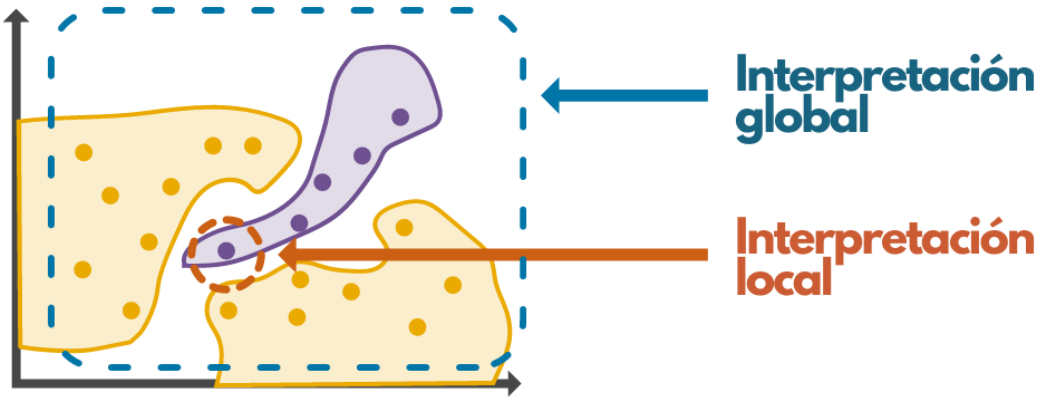


Figura 3. Enfoques de la interpretación global y local. Adaptado de [18].

En una interpretación local del modelo, la explicación está focalizada. Este enfoque no busca indicar el comportamiento del modelo en sí, sino explicar las razones por las que el modelo otorga las salidas resultantes para una entrada específica. Formar una explicación derivada de una salida segmentada a partir del espacio de la solución, dando explicaciones

a los subespacios de solución menos complejos, facilita el proceso sin dejar de lado el valor interpretativo para el usuario [17].

La segmentación en el enfoque de la interpretación local permite especificar la muestra hasta una única instancia o subconjuntos de instancias. Una representación gráfica de la focalización de cada tipo de interpretación se indica en la Figura 3. Es necesario resaltar que, en términos de fidelidad, la interpretación local no es interdependiente con la interpretación global, las características que son importantes para el modelo de manera global pueden no serlo en el contexto local y viceversa [19].

Técnicas de Inteligencia Artificial Explicada de interpretación global

En esta sección se detallan las características de las técnicas que permiten la implementación en casos concretos de la XAI enfocada en la interpretación global.

Gráfico de Dependencia Parcial

El Gráfico de Dependencia Parcial (PDP, por sus siglas en inglés) es una técnica de interpretabilidad global que indica la dependencia entre los datos de entrada y de salida obtenidos a partir de un modelo complejo de aprendizaje automático [20]. La implementación de PDP es agnóstica del modelo y su interpretación es intuitiva para el usuario [16].

PDP permite indicar el funcionamiento de un algoritmo de aprendizaje automático mediante la representación gráfica de las relaciones convencionalmente escondidas entre las entradas y salidas del modelo [21]. Estos factores son tomados como indicadores globales, es decir, no se examina una instancia en particular para indicar la influencia de características en esa muestra, sino se describe el modelo en su totalidad.

El principio detrás de PDP es bastante intuitivo, las gráficas de dependencia parcial muestran la influencia de una característica en específico, marginando los valores de todas las demás características de entrada. Esta técnica permuta los valores de la característica objetivo para obtener un valor indicativo de la influencia sobre las predicciones ejecutadas por el modelo de caja negra [21]. Un ejemplo de PDP con el dataset IRIS [22] se muestra en la Figura 4, donde se grafica la relación entre la variable predictora y los valores predichos; se puede deducir, por ejemplo, que la probabilidad de que una especia sea “virgínica” aumenta con la longitud de los pétalos de las muestras.

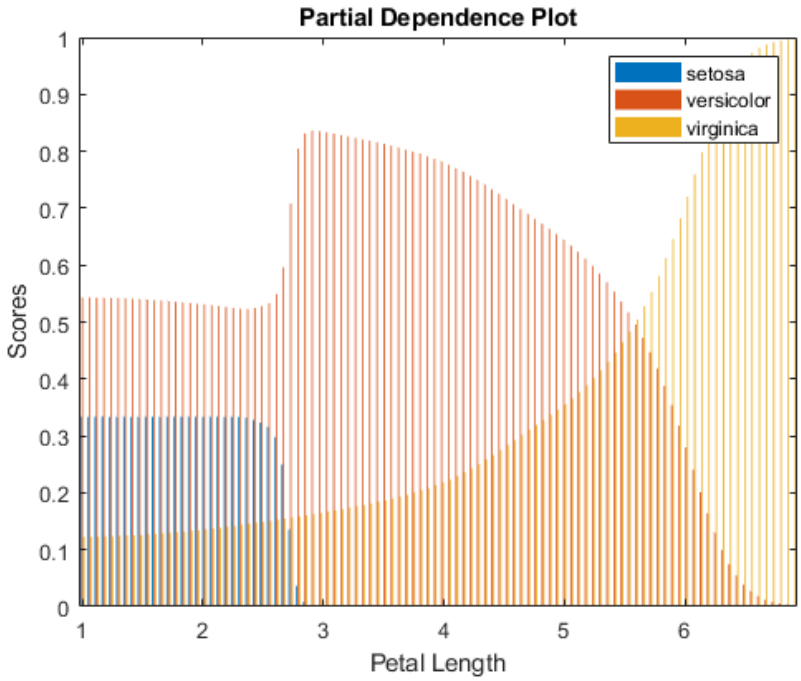


Figura 4. Ejemplo de un Gráfico de Dependencia Parcial con el dataset IRIS [23].

Existe una desventaja notable en la implementación de PDP respecto a las características que se pretenden explicar, pues esta técnica asume que cada una de las características es independiente de las demás [16]. Evidentemente, este factor limita la representación de casos existentes en la realidad, recordando la naturaleza de los datos, donde sus características suelen estar correlacionadas [24].

Efectos Acumulados Locales

A partir de las limitaciones de PDP, surge la técnica correspondiente a los Efectos Acumulados Locales (ALE, por sus siglas en inglés) [25]. ALE tiene la particularidad de que el cálculo de la influencia de una característica en el resultado de un modelo de caja negra no se da a partir de los promedios de las predicciones, sino de la diferencia entre estas.

ALE presenta la variación de las predicciones en pequeños intervalos: para las instancias de datos en un intervalo, se calcula la diferencia en la predicción cuando son reemplazadas aquellas características del límite superior e inferior del intervalo; las variaciones son acumuladas y se resta una constante para rectificar los datos, dando como resultado la curva ALE [25].

Técnicas de Inteligencia Artificial Explicada de interpretación local

En esta sección se detallan las características de las técnicas que permiten la implementación en casos concretos de la XAI enfocada en la interpretación local.

Explicaciones de Aditivos Shapley

La técnica de explicaciones de aditivos Shapley, comúnmente llamada SHAP, hace uso de los valores Shapley [26] para formular explicaciones de interpretación local agnósticas del modelo. Esta técnica fue presentada por Lundberg y Lee [27] en 2017, donde se tomó como base el fundamento detrás de los valores Shapley, el cual se profundiza en esta sección.

El teorema que describe los valores Shapley se describe a continuación [26]:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Ecuación 1. Teorema de la obtención de valores Shapley.

Este teorema permite obtener el valor Shapley ϕ para una característica específica i, en el caso de tener datos tabulados como entrada. Esta característica representaría una determinada columna. La función compleja del modelo de caja negra f hace de parámetro de entrada junto con una tupla de datos x, pues la interpretación es local.

El valor Shapley empieza con la iteración de todas las posibles combinaciones de las características que intervienen en el modelo y se relacionan con la característica específica i de la que deseamos obtener el valor Shapley ($z' \subseteq x'$).

En el segundo término del teorema, obtenemos la contribución que realiza la característica i al modelo complejo de caja negra. Para esto, evaluamos la función del modelo *f* con la característica i en las combinaciones z' posibles: $f_x(z')$, menos la salida del modelo excluyendo dicha característica: $f_x(z' \setminus i)$. Por su parte, el primer término representa la combinación que permite ejecutar la operación anterior con cada subconjunto de características y, a la vez, se pondera en función del número total de características existentes M en el conjunto de datos original.

Este principio es utilizado por SHAP para explicar las decisiones tomadas por el algoritmo de caja negra calculando la contribución que realiza cada característica en la predicción o clasificación a partir de un valor de entrada específico [16]. Esta técnica permite obtener explicaciones con modelos que utilizan valores de entrada de tipo texto o imagen, además de datos tabulares, como se indica en la Figura 5.

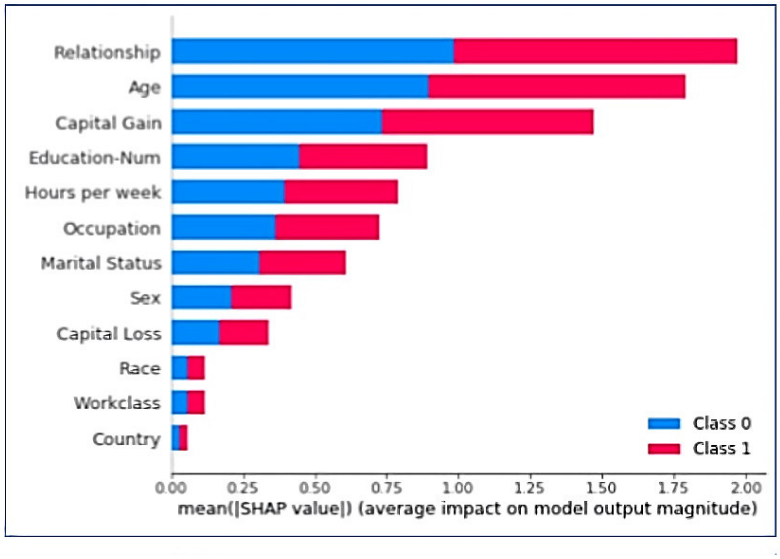


Figura 5. Ejemplo de salida para datos tabulares de la técnica SHAP [28].

Explicaciones locales interpretables agnósticas del modelo

Las explicaciones locales interpretables agnósticas del modelo (LIME, por sus siglas en inglés) denotan una propuesta relativamente innovadora que ha ganado popularidad en años recientes. Fue propuesta por Ribeiro et al. [19] en 2016, en donde se la define como "una técnica de explicación novedosa que explica las predicciones de cualquier clasificador de una manera interpretable y fiel, con el aprendizaje de un modelo localmente interpretable sobre la predicción". En términos prácticos, LIME es una técnica que otorga explicaciones de las decisiones del modelo de aprendizaje automático siendo totalmente agnóstica de este, con la particularidad de tomar como entrada una instancia específica para dar explicaciones locales [29].

En su implementación práctica, las explicaciones corresponden al tipo de dato que se utiliza para el entrenamiento del modelo de AI: en el caso de datos tabulares, LIME realiza una ponderación de las características (columnas del dataset) más y menos influyentes en las decisiones del modelo. En la Figura 6 se puede observar un ejemplo de la salida de LIME para estos casos.

Cuando los datos de entrada corresponden a segmentos de texto para análisis, LIME indica en sus explicaciones la presencia o ausencia de palabras con un porcentaje de influencia para la clasificación del modelo, un ejemplo se muestra en la Figura 7.

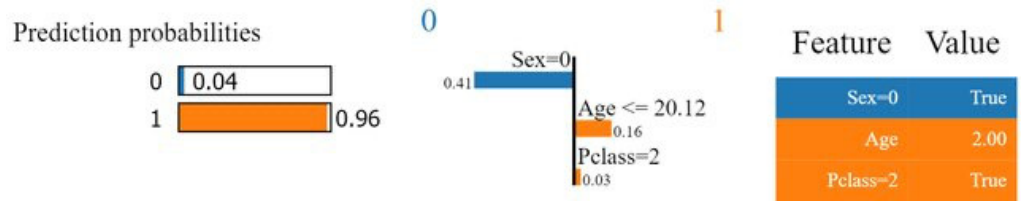


Figura 6. Ejemplo de las explicaciones de la técnica LIME para datos tabulados [30].

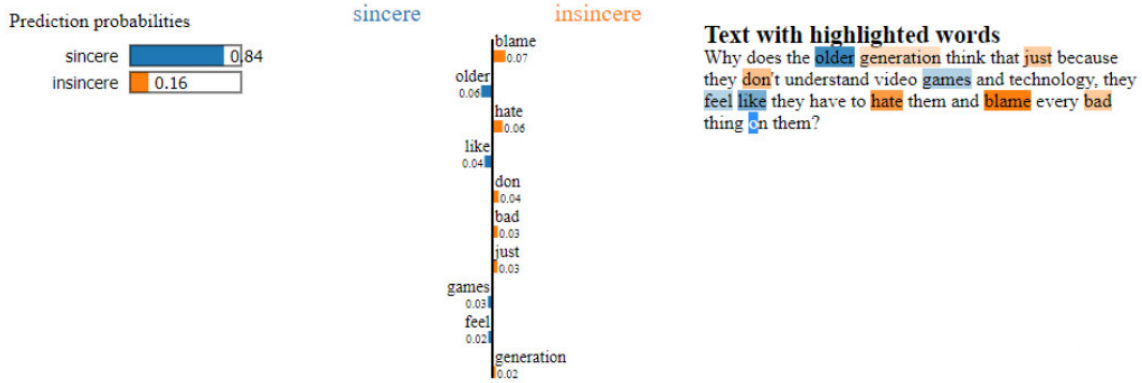
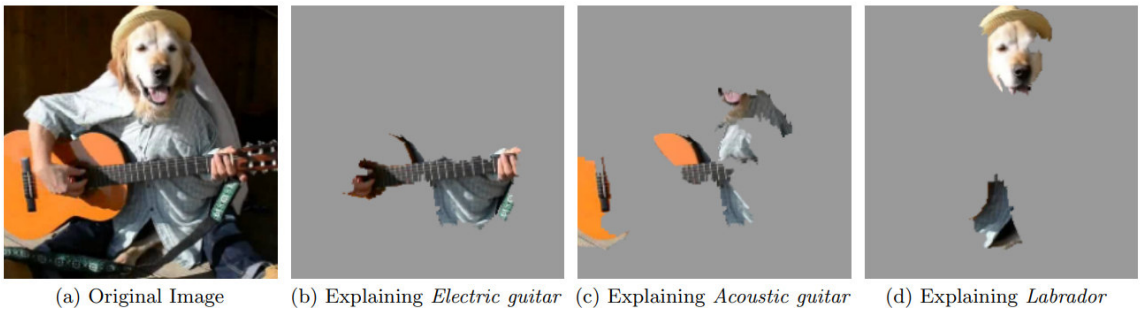


Figura 7. Ejemplo de las explicaciones de la técnica LIME para texto [31].

Para los modelos de clasificación de imágenes, la explicación se representa mediante la presencia o ausencia de superpíxeles, lo que produce segmentos de imagen interpretables para el usuario tal como se muestra en la Figura 8.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figura 8. Ejemplo de las explicaciones de la técnica LIME para imágenes [19].

Independientemente del tipo de datos de entrada que reciba el modelo de aprendizaje automático, el principio explicativo de LIME es el mismo. Esta técnica genera un nuevo conjunto de datos que consta de perturbaciones de los datos de entrada originales y las propias predicciones realizadas por el modelo de caja negra [16]. A partir de este punto, LIME entrena un modelo interpretable que se pondera mediante la proximidad de las instancias muestreadas con la instancia que nosotros hemos escogido para ser explicada, recordando la focalización al ser una técnica de interpretación local.

El problema de optimización llevado a cabo por LIME para otorgar explicaciones agnósticas del modelo se expresa a continuación [19]:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Ecuación 2. Obtención de explicaciones con LIME.

Para establecer la aproximación descrita anteriormente, tomamos como entrada a la instancia específica que pretende ser explicada. Por cuestiones de ejemplificación, se considera que el tipo de dato de entrada para este problema se encuentra en formato tabular, por lo tanto, la entrada x es una instancia con una serie de características de este tipo.

En esta ecuación, se parte por el argumento de minimización de la función, en donde un modelo interpretable local simple g pertenece a un conjunto de modelos interpretables G como pueden ser modelos lineales o árboles de decisión [32]. El primer término en la función de optimización $\mathcal{L}(f, g, \pi_x)$ representa la minimización de pérdida $\mathcal{L}$ (como el error mínimo cuadrado [33], por ejemplo); en la cual, el modelo interpretable local g intenta aproximarse a un modelo complejo de caja negra f, como una red neuronal [34], dentro de una localidad específica alrededor de los datos originales de entrada x que se definen por la medida de proximidad π_x.

El segundo término de la función trata la complejidad del modelo explicativo y se denota por $\Omega(g)$, hace la función de indicador de complejidad para el modelo interpretable local g e indica la importancia de su minimización, pues LIME se encarga de reducir la complejidad de la función de pérdida, pero es el usuario quien determina la complejidad del modelo local. Por ejemplo, si el modelo interpretable es intrínsecamente una regresión lineal, la complejidad puede estar dada por las ponderaciones que no llegan a cero [16], para reducir el $\Omega(g)$, podría ser conveniente seleccionar un número limitado de características que sean utilizadas por el modelo [35]. Este proceso se ilustra en la Figura 9.

LIME es una técnica de interpretación local relativamente fácil de implementar que se ha convertido en una poderosa herramienta para otorgar explicaciones de las decisiones tomadas por los algoritmos de caja negra [36]. Las explicaciones fácilmente entendibles para el usuario y la naturaleza agnóstica de su implementación son solo algunas de las ventajas que presenta esta técnica.

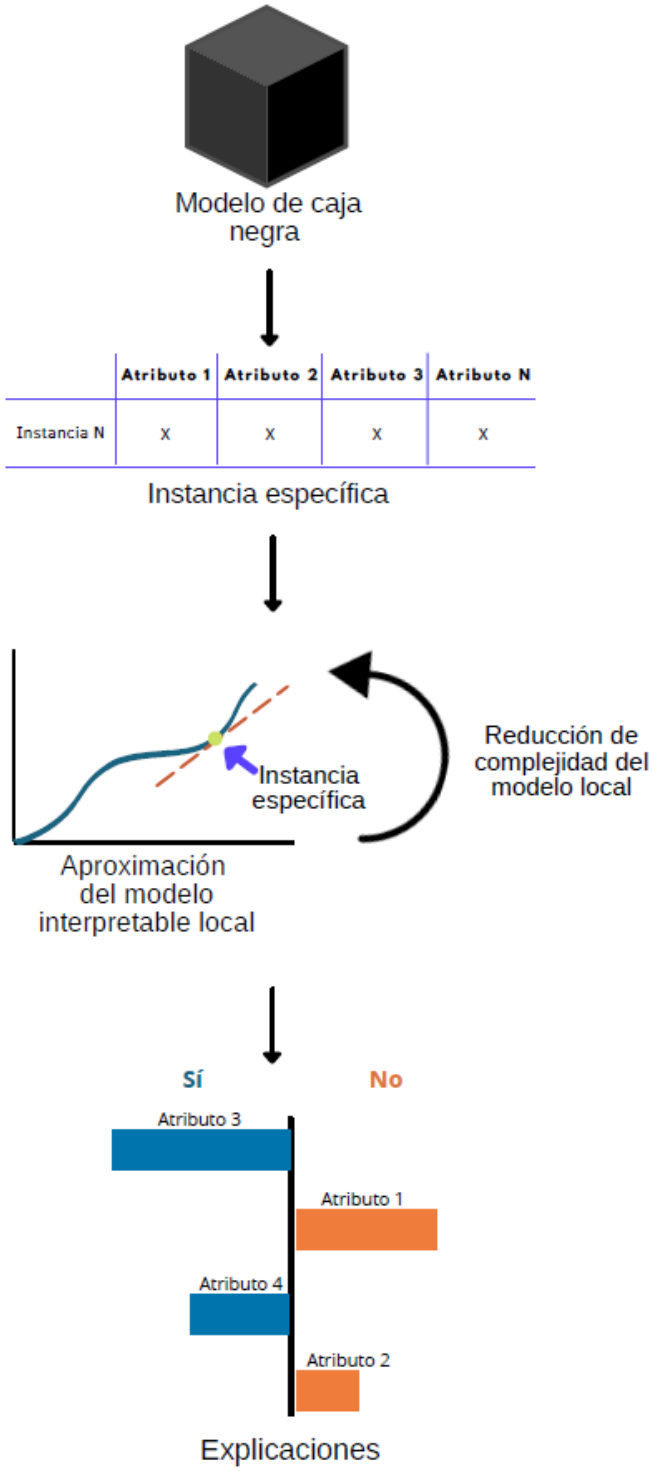


Figura 9. Funcionamiento del algoritmo de Inteligencia Artificial Explicada LIME.

1.4.4 Predicción de la salud fetal con AI

En las últimas décadas, la medicina ha encontrado en la AI un propulsor en la búsqueda de soluciones. Así, mediante algoritmos de aprendizaje automático, tareas tan específicas como el diagnóstico de pacientes se ven favorecidas por su capacidad de predicción

basada en los datos [37]. Es evidente el soporte que puede otorgar un algoritmo predictivo que no está sujeto a arbitrariedades al no disponer de la intervención subjetiva del humano.

La aplicación de los algoritmos de AI se ha hecho presente en varias ramas de la medicina, la literatura alberga estudios sobre algoritmos de aprendizaje automático que han ayudado en el diagnóstico de enfermedades específicas como Parkinson [38], diferentes manifestaciones de cáncer [39] [40], e incluso trastornos de salud mental como depresión o ansiedad clínica [41]. Esto no es más que una pequeña muestra del gran aporte de la inteligencia artificial en la salud del ser humano.

Estudios enfocados en la salud fetal

Respecto al área de obstetricia, la literatura indica estudios enfocados en la predicción de posibles complicaciones en los procesos posparto [42] [43]. También existen estudios orientados a la predicción de datos del feto como su crecimiento y desarrollo [44], o factores más específicos como la clasificación de frecuencia cardíaca fetal [45].

Existen estudios que utilizan algunos de los factores mencionados anteriormente con la finalidad de determinar la salud de un feto [46]. Su implementación suele depender del algoritmo de aprendizaje automático utilizado y el conjunto de datos aprovechable para el análisis. En ciertos casos, los datos suelen ser recolectados específicamente para el modelo a implementar [47], en otros se utilizan datos existentes que poseen un propósito general [48].

Cardiotocografía

Una de las técnicas aprovechadas tanto por los profesionales de la salud, así como por los especialistas en AI, es la Cardiotocografía (CTG). Esta técnica permite monitorizar la frecuencia cardíaca fetal y las contracciones uterinas de la madre durante las últimas etapas de gestación [49]. Es particularmente esencial en el contexto de embarazos de alto riesgo y suele presentarse en forma de examen médico. Su principal objetivo es permitir que un profesional de la salud identifique situaciones riesgosas para el bebé en etapas previas al nacimiento.

Para llevar a cabo un examen cardiotocográfico es común utilizar un dispositivo denominado cardiotocógrafo o monitor electrónico fetal. Este dispositivo capta continuamente la frecuencia cardíaca fetal y las contracciones uterinas, mientras las registra en forma de histogramas de datos que representan el resultado propiamente dicho del examen médico [50].

Las pruebas pueden realizarse por métodos internos o externos que varían en el nivel invasivo resultante para la paciente. En las pruebas externas, un cinturón provisto de dos sensores de medición se coloca alrededor del vientre de la mujer embarazada, un sensor registra los latidos del corazón del bebé y el segundo registra la intensidad y duración de las contracciones de la madre; los sensores de medición funcionan con ondas ultrasónicas inofensivas tanto para la madre como para el niño [51]. En las pruebas internas, se coloca un catéter en el útero después de que se haya producido una cantidad específica de dilatación. Esta última resulta ser la opción más precisa, pero también la más invasiva.

Interpretación de una cardiotocografía

En ambos casos, el resultado será interpretado por los especialistas en el área de obstetricia mediante el análisis de ciertos parámetros específicos. En la Figura 10 se muestra un ejemplo de una cardiotocografía obtenida a través de un monitor electrónico fetal. Si bien los elementos de la CTG no son fácilmente distinguibles, es posible diferenciar los registros de los latidos del corazón del feto y las contracciones uterinas de la madre. Generalmente, el monitor fetal es capaz de otorgar información adicional sobre lo acontecido durante el examen médico como los movimientos del feto, resaltando aquellos que fueron percibidos por la madre [50].

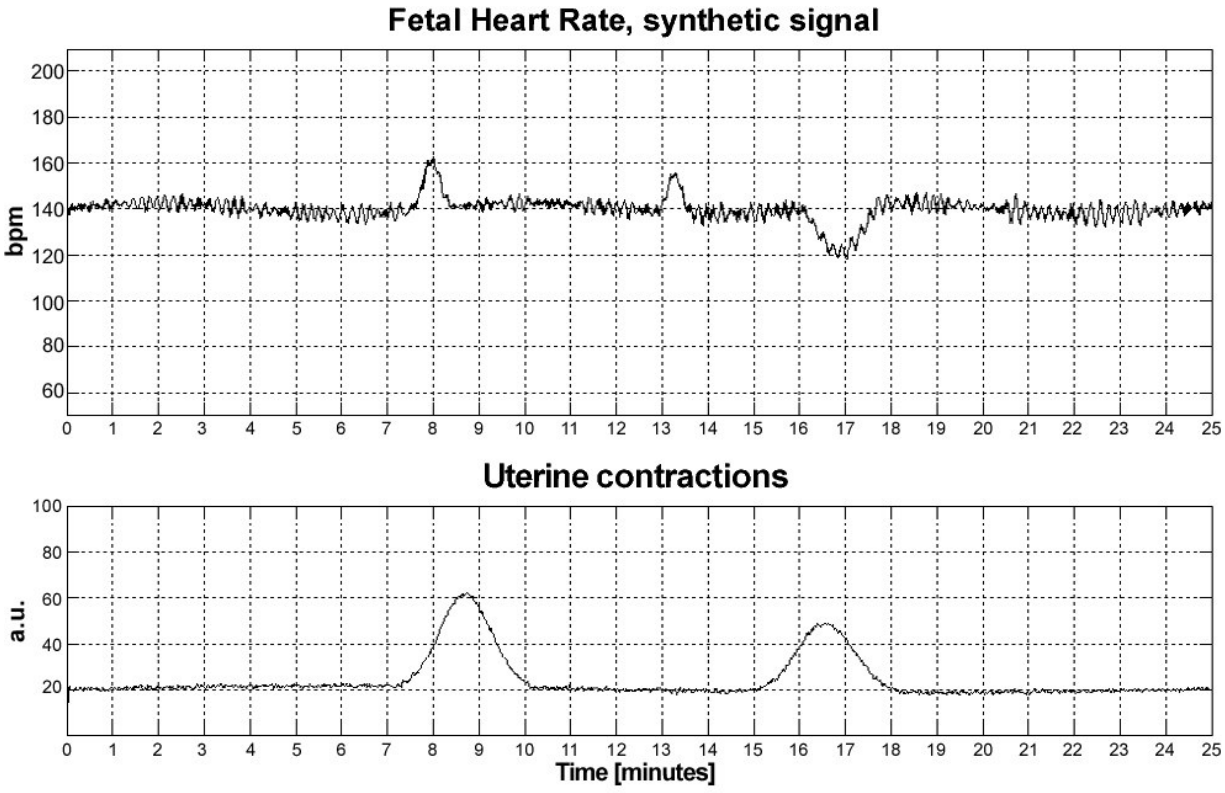


Figura 10. Ejemplo de cardiotocografía extraída con un monitor electrónico fetal [52].

2 METODOLOGÍA

La metodología para minería de datos Cross-Industry Standard Process for Data Mining (CRISP-DM) [53] guía el desarrollo de este trabajo de titulación debido a su versatilidad en la implementación de proyectos de análisis de datos y su enfoque en el entendimiento del negocio.

El objetivo principal de CRISP-DM es proporcionar un flujo de proceso estandarizado independiente de la industria o del software que se va a utilizar y de la minería de datos que se va a ejecutar [54]. Por esta razón, CRISP-DM es una de las metodologías más utilizadas en proyectos de ciencia de datos y AI [55]. Adicionalmente, la naturaleza flexible de CRISP-DM permite que sus fases no sigan una secuencia estricta, es común observar casos donde es necesario regresar a fases anteriores del modelo de proceso.

En la Figura 11 se ilustra el ciclo correspondiente a la metodología CRISP-DM. Cada una de las fases tiene un propósito específico y se detalla a continuación [53]:

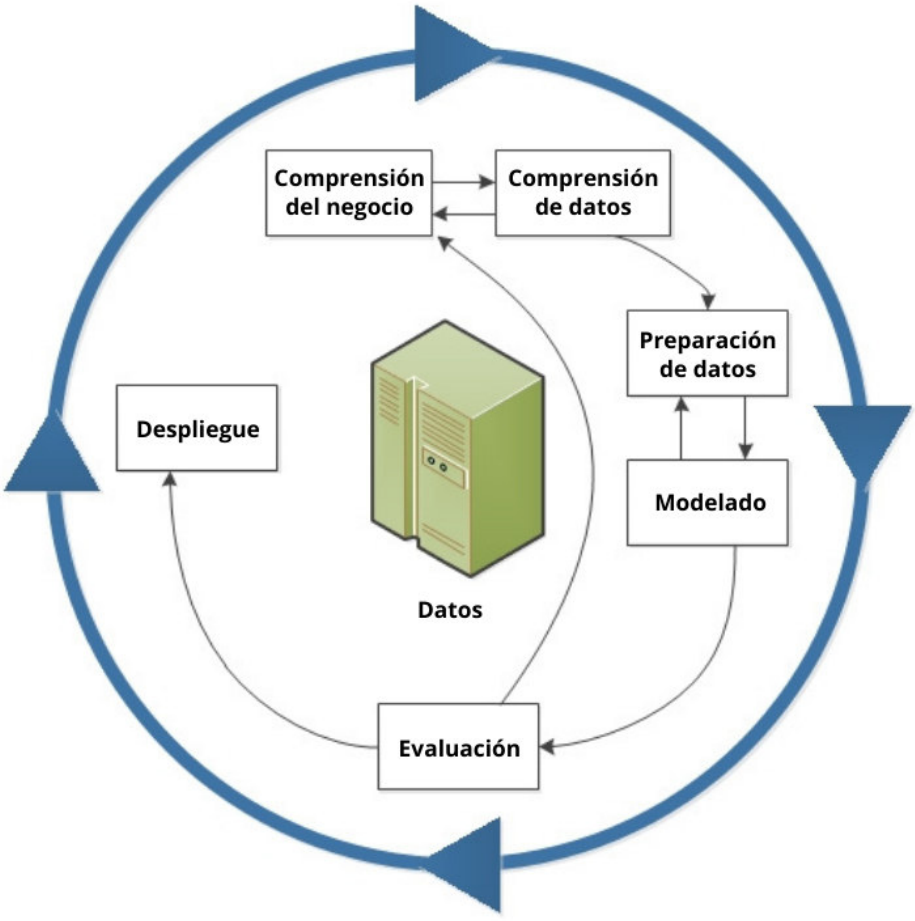


Figura 11. Fases de CRISP-DM. Adaptado de [53].

1. **Comprensión del negocio:**

La fase de comprensión del negocio consiste en establecer objetivos y requisitos específicos que guiarán el proyecto en el contexto de la minería de datos. Es esencial evaluar la problemática mientras se tienen presentes las necesidades del cliente o el público beneficiado con el proyecto. Esta fase puede llegar a ser la más importante del proceso, pues constituirá la base para las etapas siguientes.

2. **Comprensión de datos:**

Como parte de la comprensión de datos, se intenta obtener una descripción detallada de los datos disponibles y entender la naturaleza de las características que componen el conjunto de entrada sin un previo refinamiento. Adicionalmente, es necesario evaluar la calidad de los datos en función de su capacidad para cumplir los objetivos del negocio descritos en la fase anterior.

3. **Preparación de datos:**

La preparación de datos se utiliza para crear un conjunto de datos final que forma la base para la etapa de modelado. Comúnmente, suele representar la fase más larga del proyecto, pues consiste en corregir, reemplazar o eliminar valores erróneos o faltantes como parte de una limpieza minuciosa de cada uno de los registros en el conjunto de datos. La correcta ejecución de esta fase puede suponer el éxito de la fase de modelado en cuanto a temas de precisión y exactitud.

4. **Modelado:**

Como regla general, se pueden utilizar varias técnicas de modelado de minería de datos para un problema determinado. Algunas técnicas imponen requerimientos específicos a la estructura de los datos de entrada, esto puede implicar la necesidad de dar un paso atrás hacia la fase de preparación de datos. Entre las actividades típicas de esta fase está la selección de técnicas de modelado, la elaboración de un plan de pruebas para verificar la precisión del modelo y la implementación.

5. **Evaluación:**

El proyecto de minería de datos se evalúa retrospectivamente, en esta fase se verifica que el modelo realmente cumple con los objetivos del proyecto de minería de datos. Si los objetivos no se pudieron alcanzar, la fase puede volver a ejecutarse. Generalmente, esta evaluación se lleva a cabo de manera cualitativa, pues las evaluaciones cuantitativas de

los modelos se ejecutan en la fase anterior. Finalmente, se determina si los atributos del modelo desarrollado podrían ser utilizados para futuros proyectos de minería de datos.

6. **Despliegue:**

En la fase del despliegue los conocimientos adquiridos se organizan y presentan de tal manera que el usuario tenga la oportunidad de utilizarlos. Esto incluye una posible estrategia de implementación, seguimiento de la validez de los modelos, un sistema interactivo, entre otras opciones.

Como elementos claves de CRISP-DM, aplicados al problema planteado, se pueden identificar dos. Por un lado, la orientación en el entendimiento del negocio, pues es esencial adquirir el conocimiento base necesario para desarrollar un proyecto de investigación en áreas de riguroso análisis como la medicina. Por otro lado, el enfoque en la evaluación de modelos otorga la oportunidad de identificar los algoritmos más precisos en cuanto a predicción y explicación, para este último factor se planea identificar el grado de utilidad real del algoritmo de XAI mediante una evaluación cualitativa con médicos especializados en el área.

A continuación, se describe la aplicación de cada una de las fases de la metodología CRISP-DM al problema planteado:

2.1 Comprensión del negocio

Debido al enfoque investigativo de este trabajo de titulación, se plantea una revisión sistemática de literatura para apoyar la comprensión del negocio en una primera instancia y determinar la cuantía investigativa existente hasta el momento.

En términos prácticos, en esta sección se determinan los objetivos de negocio identificando las necesidades de los profesionales médicos para mejorar el diagnóstico de la salud fetal. Se lleva a cabo una revisión sistemática de literatura que permita conocer las aplicaciones de la XAI en la medicina y, a partir de este punto, profundizar en áreas referentes a la salud fetal.

2.1.1 Revisión Sistemática de Literatura

Esta sección basa su desarrollo en la metodología Kitchenham para la elaboración de Revisiones Sistemáticas de Literatura orientadas a la Ingeniería de Software [56]. Este enfoque se utiliza para garantizar una investigación rigurosa, metódica y confiable que permita obtener resultados de interés a partir de interrogantes específicas relacionadas con el objetivo de la investigación.

Preguntas de investigación para SLR

Las preguntas de investigación que guían esta SLR son:

1. ¿Cuáles son las áreas de la medicina en las que se han aplicado algoritmos de Inteligencia Artificial Explicable?
2. ¿Qué problemas de la medicina son resueltos por la XAI?
3. ¿Qué metodologías (algoritmos de AI / XAI) llevan a cabo estos estudios?
4. ¿Cuál es la naturaleza de los conjuntos de datos utilizados en estos estudios?
5. ¿Qué técnicas de validación son utilizadas para evaluar los modelos de AI / XAI utilizadas en estos estudios?

Con las preguntas de investigación como base, es posible iniciar el proceso de implementación de la SLR. Para esto, Kitchenham proporciona tres fases intuitivamente delimitadas [56]:

1. **Fase de búsqueda**

En esta fase se definen los parámetros iniciales para obtener el paquete de literatura preliminar. Las bases de datos científicas que fueron tomadas como fuente bibliográfica para extraer la información son: IEEE Xplore y ACM Digital Library. Esta decisión se basa en la calidad de la información que puede ser encontrada en estas bases de datos, su popularidad y en los métodos de búsqueda disponibles.

Palabras clave

Como punto de partida se definieron algunas palabras clave para la revisión sistemática tomando como base las preguntas de investigación:

- Inteligencia Artificial
- Aprendizaje automático
- Inteligencia Artificial Explicada
- Interpretable
- Explicable
- Medicina
- Médico

Cadena de búsqueda

El siguiente elemento definido en esta fase corresponde a la cadena de búsqueda que será ejecutada en la sección posterior. Su construcción se basa en las palabras clave y pretenden obtener literatura que responda a las preguntas de investigación. Los elementos de la cadena de búsqueda están en inglés por cuestiones concernientes a los sistemas de búsqueda de las bases de datos científicas. La cadena se muestra a continuación:

SS: *En abstract: (Explained Artificial Intelligence OR Explainable Artificial Intelligence OR XAI OR Interpretable AI) AND En toda la metadata: (Medicine OR Medical)*

Como se observa, la cadena de búsqueda está dividida para ser condicionada mediante los criterios que ofrecen las bases de datos científicas. Específicamente, los términos relacionados a la XAI serán explorados en la sección del resumen o abstract de los artículos científicos; mientras tanto, los términos relacionados a la medicina serán indagados en todo el contenido disponible (all metadata).

Estas cadenas se modificaron ligeramente para adaptarse al patrón de búsqueda de cada base de datos. Se utilizaron las siguientes variaciones:

IEEE Xplore: *("Abstract": Explained Artificial Intelligence OR "Abstract": Explainable Artificial Intelligence OR "Abstract": XAI OR "Abstract": Interpretable AI) AND ("All Metadata": Medicine OR "All Metadata": Medical)*

ACM DL: *[[Abstract: "Explained Artificial Intelligence"] OR [Abstract: "Explainable Artificial Intelligence"] OR [Abstract: XAI] OR [Abstract: "Interpretable AI"]] AND [[All: Medicine] OR [All: Medical]]*

Criterios de inclusión y exclusión

Se definen los criterios de inclusión y exclusión que permitirán efectuar un filtrado inicial en las siguientes fases. Estos criterios se enumeran a continuación:

Inclusión:

- Artículos publicados en los últimos 5 años.
- Estudios experimentales.
- Artículos escritos en inglés.
- Estudios que hayan sido publicados en revistas y conferencias.

Exclusión:

- Estudios incompletos o en proceso.
- Artículos con un número de páginas menor a 3.

2. Fase de ejecución

La fase de ejecución tiene un principio bastante intuitivo, consiste en la ejecución de la cadena de búsqueda en las bases de datos científicas definidas en la fase anterior, de esta forma se obtiene el conjunto inicial de literatura que no presenta ningún refinamiento adicional a los criterios de inclusión y exclusión. Posteriormente, el conjunto inicial es filtrado en función de su contenido hasta obtener una lista de estudios de calidad respecto a lo cuestionado en las preguntas de investigación.

La cadena de búsqueda fue ejecutada el 22 de noviembre de 2021 y se obtuvieron los resultados descritos en la Tabla 1.

Tabla 1. Número de artículos resultante de la ejecución de la cadena de búsqueda.

Base de datos	Resultados
ACM	92
IEEE	32
Total:	124

El conjunto inicial obtenido a partir de la ejecución de las cadenas de búsqueda es de 124 elementos; se ha constatado que este grupo no contenga artículos duplicados. A partir de este punto inicia el filtrado de estudios acorde con su contenido. Este es un proceso manual en el que el investigador evalúa la calidad del estudio y su potencial para responder las preguntas de investigación.

El primer filtro corresponde al análisis de los títulos y los abstracts de los estudios resultantes. Los criterios analizados en esta etapa corresponden principalmente a la inclusión de estudios en donde se apliquen algoritmos de XAI en la medicina, independientemente de los resultados o las validaciones efectuadas. El número de artículos descartados con este filtro es de 77, el número de artículos resultante es de 47.

Evidentemente, un filtrado a partir de los títulos y abstracts no es suficiente para garantizar la calidad de los estudios. Por esta razón, se llevó a cabo un segundo filtrado concerniente al análisis minucioso de cada uno de los artículos, priorizando a aquellos que podrían responder las preguntas de investigación. Varios trabajos fueron descartados debido al reducido alcance de su estudio que, en ciertos casos, limitaba la extracción de información.

El número de artículos descartados con este filtro es de 20, el número de artículos resultante es de 27.

Este conjunto de artículos representa la literatura definitiva de la que se extraerá la información para sintetizar la evidencia científica concerniente a las aplicaciones de la XAI en la medicina. El proceso de ejecución se resume en la Figura 12.

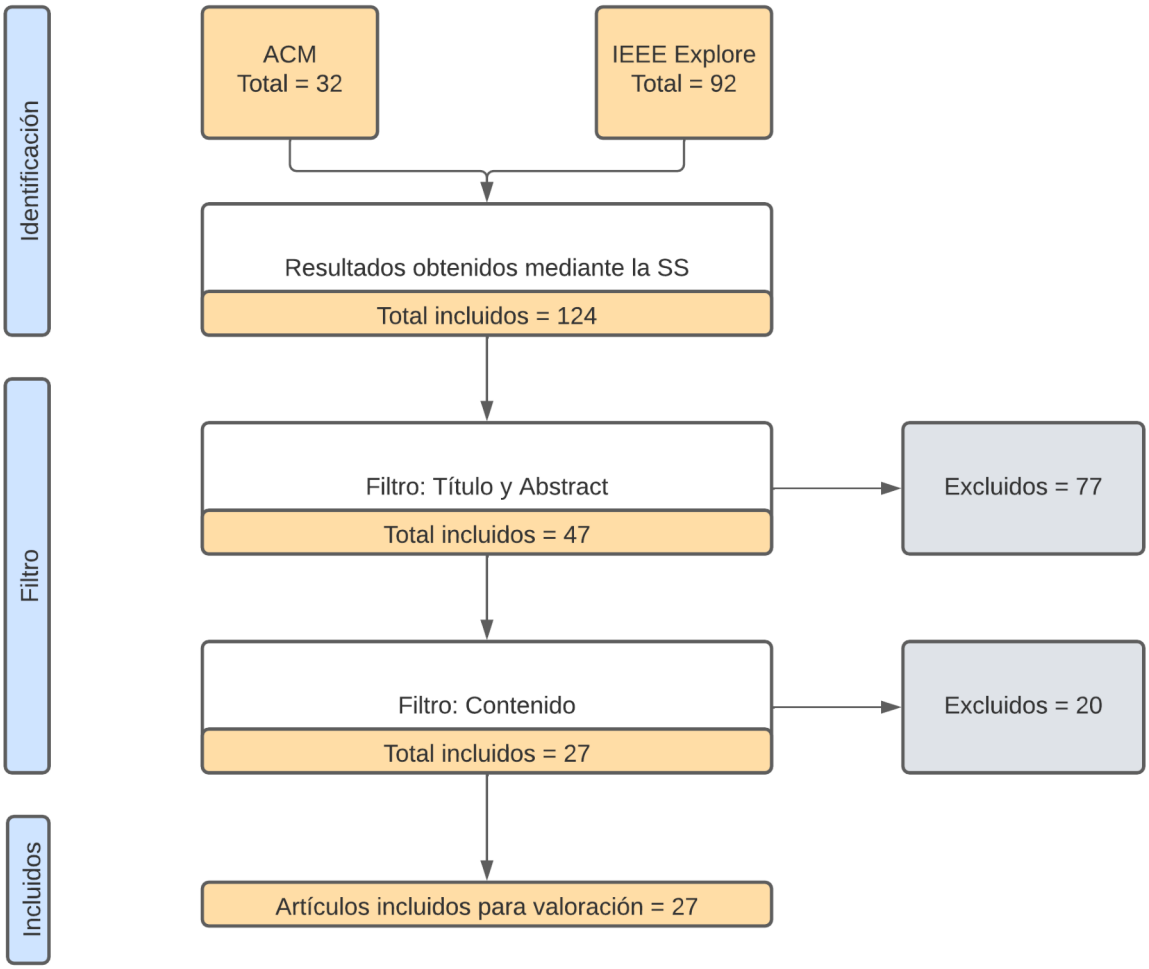


Figura 12. Diagrama de flujo de la fase de ejecución de la SLR.

3. Fase de extracción de información

Tomando como punto de partida el conjunto de artículos resultante de la sección anterior, es necesario extraer la información relevante de cada uno de los estudios que fueron seleccionados sistemáticamente.

Criterios de extracción

El proceso de extracción de información se ha realizado con la intención de responder a las preguntas de investigación planteadas en esta revisión sistemática, por lo tanto, los

datos extraídos responden directamente a estas interrogantes. Cada elemento recabado por el investigador resulta de un análisis minucioso del estudio científico; estos elementos se listan a continuación:

- Área de la medicina en la que se enfoca el estudio.
- Problema que se busca resolver.
- Solución o propuesta al problema.
- Algoritmos de AI y XAI implementados en la solución.
- Naturaleza del conjunto de datos utilizado (existente o recolectado).
- Métodos de validación de los algoritmos implementados.

La extracción de los datos con los 27 artículos finales se puede observar en el Anexo 2.

2.1.2 Resultados de la Revisión Sistemática de Literatura

La sección referente a la fase de extracción de información de la SLR proporcionó un conjunto de datos extraídos a partir de la serie de artículos científicos seleccionados sistemáticamente. La presente sección, por su parte, brinda un análisis de este conjunto de datos con la intención de discutir los puntos relevantes que responden a las preguntas de investigación planteadas.

XAI implementada en medicina

Como se ha mencionado en secciones anteriores, la AI es aplicable a muchos campos de la medicina; sin embargo, el nacimiento reciente de la XAI y su naturaleza innovadora provoca que este tipo de algoritmos no contenga abundante literatura. A través de la respuesta a la primera pregunta de investigación es posible corroborar o refutar el punto anterior.

Cada una de las áreas de la medicina en las que se aplicó un algoritmo de XAI en los 27 artículos finales se indica en la Tabla 2, con orden descendente en función del número de estudios de cada área. Esta lista responde a la primera pregunta de investigación.

Son 14 los campos en los que se han aplicado algoritmos de XAI en los estudios resultantes de la SLR. Encabezando la lista con 6 estudios se encuentra la neurología, área considerada de interés en el ámbito científico; siguiendo el orden se halla la cardiología con 3 estudios, al igual que la oncología y la gestión hospitalaria. La lista se completa con

otros campos recurrentes en la investigación médica, dando un panorama general de las aplicaciones que pueden ser llevadas a cabo por la XAI en la medicina.

Tabla 2. Áreas de la medicina en las que se aplicaron algoritmos de XAI.

N.º	Área de la medicina	Estudios	Total
1	Neurología	[57], [58], [59], [60], [61], [62]	6
2	Cardiología	[63], [64], [65]	3
3	Oncología	[66], [67], [68]	3
4	Gestión hospitalaria	[69], [70], [71]	3
5	Procesamiento de imágenes biomédicas	[72], [68]	2
6	Epidemiología	[73], [74]	2
7	Educación quirúrgica	[75]	1
8	Medicina general	[76]	1
9	Reumatología	[77]	1
10	Endocrinología	[78]	1
11	Neumología	[79]	1
12	Odontología	[80]	1
13	Oftalmología	[81]	1
14	Dermatología	[82]	1

El factor que advierte el interés de esta investigación corresponde a la falta de literatura en el área de obstetricia. La revisión sistemática de literatura llevada a cabo en este trabajo de titulación revela la inexistencia de estudios que indiquen la aplicación de algoritmos de XAI en la obstetricia y, por lo tanto, en el apoyo a la toma decisiones médicas en la salud fetal.

Problemas médicos resueltos por la XAI

El Anexo 2 resultante de la fase de extracción de información enumera una serie de problemáticas que pretenden ser resueltas por la XAI. Para cada problema, el estudio plantea una solución ligada principalmente a la capacidad explicativa de los algoritmos de XAI, por lo tanto, esta es una muestra de las oportunidades que brinda el hecho de conocer el funcionamiento de los algoritmos de caja negra e interpretar sus decisiones.

En respuesta a la segunda pregunta de investigación es posible resaltar algunos estudios para ejemplificar las situaciones médicas en las que la XAI cumple un rol resolutivo. Tal es el caso del estudio de Davagdorj et al. [76], en donde advierte que el rápido aumento de las enfermedades no transmisibles (ENT) se ha convertido en un grave problema de salud en el mundo.

Para los autores, los sistemas basados en AI desarrollados en los últimos años son poco interpretables para los profesionales de la salud por su naturaleza de estilo de caja negra. Con la intención de abordar este problema, se propone una red neuronal profunda (DNN) basada en DeepSHAP equipada con una técnica de selección de características para construir un sistema de soporte de decisiones preciso y explicable.

Adicionalmente, la revisión sistemática arrojó estudios que proponen soluciones ad-hoc, es decir, el diseño y la implementación del modelo explicativo se efectúa según la problemática planteada. Ejemplo de ello es el estudio de Karim et al. [77], quienes mencionan la existencia de dificultades en la detección de osteoporosis por imágenes médicas, lo que representa un desafío enorme en su cuantificación temprana. Si bien los algoritmos de AI convencionales cubren la labor de detección; para los autores, la falta de interpretabilidad representa conflictos éticos y legales, además de la incapacidad de emitir decisiones sobre los diagnósticos médicos. En este estudio se propone un método novedoso y explicable para el diagnóstico de la osteoartritis de rodilla basado en radiografías y resonancia magnética, este método lleva por nombre DeepKneeExplainer.

Esta es una pequeña muestra de la literatura alrededor de las problemáticas resueltas por la XAI.

Metodologías implementadas

Desde los orígenes de la AI hasta la actualidad, numerosas técnicas y algoritmos han sido creados con la intención de resolver diferentes tipos de problemas. En la Tabla 3 se observan las técnicas de AI que se incluyeron en la mayor cantidad artículos en la SLR. El listado completo se encuentra en el Anexo 3.

Tabla 3. Principales técnicas de AI implementadas en los estudios resultantes de la SLR.

N.º	Algoritmo de AI	Estudios	Total
1	Random Forest	[57], [76], [70], [78], [58], [73], [65], [67], [83]	9
2	CNN	[57], [63], [72], [77], [81]	5
3	DNN	[76], [71], [61], [82]	4
4	Support Vector Machine (SVM)	[57], [76], [78], [58], [60]	5

Sin embargo, lo que sorprende en este estudio es la cantidad de técnicas y algoritmos de XAI que han sido implementados desde su reciente auge. En la Tabla 4 se muestran 31 técnicas y algoritmos de XAI implementados en la literatura obtenida con la revisión sistemática; esta lista responde a la tercera pregunta de investigación.

Tabla 4. Técnicas de XAI implementadas en los estudios resultantes de la SLR.

N.º	Técnica de XAI	Estudio	Total
1	SHAP (SHapley Additive exPlanations)	[59], [63], [65], [66], [71], [74], [80]	7
2	LIME (Local Interpretable Model-Agnostic Explanations)	[66], [74], [78], [82]	4
3	Grad-CAM++	[57], [77]	2
4	Class Activation Mapping	[68], [74]	2
5	Deep Shapley Additive Explanations (DeepSHAP)	[76]	1
6	Técnica XAI para generar un mapa de prominencia	[75]	1
7	Guided Backpropagation	[72]	1
8	Explainable Boosting Machine (EBM)	[66]	1
9	Scoped Rules	[66]	1
10	Learning Vector Quantization (LVQ) y variaciones (Interpretable)	[64]	1
11	Layer-Wise Relevance Propagation (LRP)	[77]	1
12	Fast-and-Frugal Tree	[70]	1
13	Análisis de patrón multivariante impulsado por un algoritmo genético (Fuzzy Logic).	[58]	1
14	TreeExplainer	[59]	1
15	Eli5	[65]	1
16	Direct backpropagation	[60]	1
17	FARC-HD como método de explicación	[67]	1
18	Permutation feature importance	[80]	1
19	Morris Sensitivity Analysis	[80]	1
20	TCAV	[81]	1
21	Saliency	[61]	1
22	Input x Gradient	[61]	1
23	Feature Ablation	[61]	1
24	Feature Permutation	[61]	1
25	Layer Wise Relevant Propagation	[61]	1
26	Deconvolution	[61]	1
27	Score-CAM (SCAM)	[61]	1
28	Counterfactual Generative Network	[79]	1
29	Gráfico Acíclico Dirigido(DAG)	[83]	1
30	EasyPEASI	[62]	1
31	Doctor XAI	[69]	1

Esta información indica que los algoritmos de XAI predominantemente utilizados en el área de la medicina son SHAP y LIME con 7 y 4 estudios, respectivamente. Los demás algoritmos presentes en este conjunto de investigaciones son un compendio de técnicas

que proponen entender las decisiones detrás de los algoritmos de AI o incluso mejorar estas técnicas a un nivel explicativo superior al propuesto inicialmente.

Conjuntos de datos utilizados

En respuesta a la cuarta pregunta de investigación, la Tabla 5 indica los estudios que utilizaron un conjunto de datos existente antes de la investigación y los estudios que establecieron un proceso de recolección de datos para la implementación de sus modelos de aprendizaje automático y XAI. El primer grupo prepondera con 20 artículos, mientras que los conjuntos de datos recolectados están presentes en 7 estudios.

Tabla 5. Clasificación de los estudios según el conjunto de datos utilizado.

Dataset	Estudios	Total
Existente	[75], [76], [63], [66], [73], [68], [77], [58], [78], [70], [79], [59], [65], [60], [67], [61], [82], [83], [62], [69].	20
Recolectado	[57], [72], [64], [80], [71], [81], [74]	7

Métodos y técnicas de validación de algoritmos

La quinta y última pregunta de investigación corresponde a los métodos de validación utilizados para evaluar los algoritmos de aprendizaje automático y los algoritmos de XAI. Es necesario mencionar que los métodos de evaluación utilizados para determinar la capacidad de predicción y su exactitud, por lo general, no son de utilidad para evaluar la capacidad explicativa de un algoritmo y viceversa.

El Anexo 2 presenta una columna dedicada exclusivamente a este apartado, en donde es posible notar que cerca de la totalidad de las técnicas evalúan únicamente los algoritmos de aprendizaje automático. Este resultado es un indicador de la escasez de técnicas de validación de algoritmos de XAI presentes en la literatura, más aún para las técnicas en las que participan los profesionales del campo en el que se aplica el algoritmo, en este caso, los médicos. Los dos únicos estudios que proponen una evaluación de la interpretabilidad del modelo se discuten a continuación:

Dong et al. [72] establecen una evaluación subjetiva con 10 participantes, la mitad de ellos eran expertos en el área informática en la que recae la investigación, la otra mitad no disponía de ninguna experiencia. Los participantes evaluaron los resultados de diferentes enfoques sin saber cuál era cada uno; al final, el enfoque propuesto por los autores obtuvo los puntajes más altos, evidenciando la capacidad de su propuesta para ser utilizada por usuarios con poca o mucha experiencia.

Kumar et al. [68], por su parte, realizan comparaciones de los resultados con modelos médicos (visualizaciones de retroalimentación CAM) y efectúan encuestas a expertos, así como comparaciones mediante diagnósticos manuales. Luego de esta validación, los autores garantizan que su propuesta es una solución confiable para los profesionales de la salud que requieran utilizarla.

2.1.3 Estado del arte

Este trabajo de titulación enfoca su investigación en el algoritmo de XAI denominado Local Interpretable Model-Agnostic Explanations. Con la intención de conocer el estado del conocimiento acumulado actual en esta área específica, se plantea un estado del arte tomando como entrada los estudios resultantes de la SLR que implementaron el algoritmo LIME en su metodología.

J. Duell, X. Fan, B. Burnett, G. Aarts, y S.-M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records", en *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4, doi: 10.1109/BHI50953.2021.9508618.

Este estudio plantea una evaluación comparativa de las explicaciones dadas por los métodos XAI como una extensión terciaria en el análisis de registros médicos electrónicos (HCE) complejos. En particular, se estudia la mortalidad por cáncer de pulmón con la premisa de que la predicción temprana de la mortalidad de los pacientes con cáncer de pulmón puede ayudar a identificar a aquellos que se beneficiarán de cierto tratamiento y a los que están en riesgo de recaída. Para los autores, el uso de XAI permitiría informar a los expertos en el dominio aquellas características de importancia difíciles de deducir para un humano debido a la cantidad de datos.

Se comparan las características de los HCE en términos de su importancia de predicción estimada por los modelos XAI. En este caso los algoritmos comparados son LIME, SHAP y las denominadas reglas de alcance. Para la implementación de este estudio, los autores hicieron uso de un conjunto de datos sintéticos obtenido a partir de datos anónimos de casos de cáncer.

Los resultados experimentales indican una variación en las características que se consideran importantes por los métodos de XAI estudiados. Aunque los tres métodos: SHAP, LIME y las reglas de alcance han identificado a la misma característica como la más importante para decidir la mortalidad de un paciente, difieren en la identificación de características secundarias o terciarias. Los autores indican que las ilustraciones de SHAP

aportan claridad al comunicar una explicación de un problema, proporcionando más que una simple predicción.

P. F. Khan y K. Meehan, "Diabetes prognosis using white-box machine learning framework for interpretability of results", en *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 2021, pp. 1501–1506, doi: 10.1109/CCWC51732.2021.9375927.

Este artículo presenta un estudio de caso sobre diabetes en mujeres nativas americanas Pima. Los autores afirman que la detección temprana de la diabetes puede ayudar a los médicos en el tratamiento de los pacientes para evitar o reducir los riesgos de complicaciones como infarto de miocardio, amputaciones de extremidades e insuficiencia renal. Sin embargo, la AI puede predecir la presencia de diabetes, pero no los factores que influyen en esta predicción.

Se propone la comparación de modelos de AI y, luego de seleccionar aquellos que otorgan la mayor precisión, se implementa LIME con la finalidad de dar una explicación clara a las decisiones del modelo. El estudio utiliza el conjunto de datos de diabetes Pima Indians alojado en el repositorio University of California Irvine's (UCI) Machine Learning Lab [84].

La precisión más alta se obtuvo con un modelo de caja negra de tipo Random Forest, alcanzando un porcentaje del 80,5% en la predicción. LIME, por su parte, proporcionó información sobre los factores contribuyentes para cada caso e indicó que el nivel de glucosa fue la característica más decisiva en la predicción de la presencia de diabetes.

F. Stieler, F. Rabe, y B. Bauer, "Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach", en *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1802–1809, doi: 10.1109/CVPRW53098.2021.00199.

Este estudio indica la manera en la que las explicaciones de un sistema de inteligencia artificial para el análisis de imágenes pueden ser más específicas para un dominio concreto. En este caso el dominio corresponde al diagnóstico dermatológico mediante un clasificador de imágenes de piel.

Según los autores, el uso de algoritmos de caja negra en entornos críticos como el campo médico requiere de la implementación de técnicas de XAI para su interpretación y, adicionalmente, las explicaciones deben adaptarse al problema para que sean útiles en el caso de uso específico.

El enfoque del estudio consiste en sintetizar la metodología de interpretación de LIME con la regla ABCD de dermatoscopia, un procedimiento de diagnóstico humano para distinguir lesiones cutáneas melanocíticas y no melanocíticas. Para convertir al modelo en uno específico del dominio, se modificó el algoritmo de perturbación de LIME a lo largo de las dimensiones de la regla ABCD. Con esta modificación es posible formular hipótesis sobre las predicciones del modelo de caja negra.

Para los autores, el resultado del explicador es útil inclusive sobre modelos de caja negra que no presentan una alta solidez y precisión. El modelo de XAI proporciona información relevante sobre las predicciones, lo que puede ayudar a mejorar tanto la toma de decisiones como el desarrollo y perfeccionamiento del modelo.

Q. Ye, J. Xia, y G. Yang, "Explainable AI for COVID-19 CT Classifiers: An Initial Comparison Study", en *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 521–526, doi: 10.1109/CBMS52027.2021.00103.

Este estudio tiene como objetivo construir un sistema de clasificación de imágenes de COVID-19 que incorpore algoritmos de aprendizaje profundo y de XAI, integrando su capacidad explicativa con modelos existentes como LIME y SHAP. La funcionalidad de este sistema basado en datos es la de proporcionar resultados de clasificación precisos y explicables para un modelo de aprendizaje profundo.

La IA puede respaldar la evaluación rápida de las tomografías computarizadas (TC) para diferenciar la presencia de COVID-19 de otras enfermedades pulmonares. Sin embargo, no está claro cómo estos algoritmos de aprendizaje profundo toman tal decisión y cuáles son las características más influyentes que podrían servir para emitir juicios o diagnósticos. Estas razones justifican la creación de un modelo de XAI.

Las imágenes del conjunto de datos de entrada, recolectado de cuatro hospitales de China, fueron divididas en superpíxeles, el estudio adoptó el método LIME y se calcularon los valores de Sharply con el fin de interpretar la contribución individual de cada superpíxel. En términos prácticos, estos algoritmos tienen la labor de determinar qué parte de las tomografías computarizadas contiene las áreas de lesión más predictivas, lo que lleva al modelo de caja negra a tomar tales decisiones.

Acorde a los resultados, el método propuesto es aplicable como una valiosa herramienta de diagnóstico auxiliar para los radiólogos, pues les permitiría realizar juicios basados en los resultados proporcionados por los mapas de activación resultantes. Sin embargo, en el

estudio no se llevó a cabo una evaluación participativa con médicos especialistas en el área para verificar esta afirmación.

2.1.4 Objetivos y necesidades identificados

Los estudios analizados en esta sección permitieron identificar ciertos aspectos relevantes para la identificación de necesidades y objetivos del problema. Una constante que se repite es la búsqueda por mejorar la toma de decisiones respecto al dominio del diagnóstico médico o la identificación de características influyentes en la detección temprana de patologías.

Algunos estudios buscan conocer el mejor algoritmo de aprendizaje automático, para lo que plantean una comparativa donde seleccionan al que otorga los mejores resultados basándose en métricas cuantitativas; a partir de este punto, implementan el algoritmo de XAI para encontrar las variables más influyentes que les permitan tomar las decisiones correctas.

Esto sería, en resumen, los objetivos y necesidades que han sido rescatados de la investigación realizada. A partir de este punto, se desarrolla la metodología planteada, asentando estos aspectos al campo de la obstetricia y la clasificación de la salud fetal, específicamente.

2.2 Comprensión de los datos

En esta fase se llevará a cabo un análisis preliminar de los datos que servirán de entrada para los algoritmos de AI que se considerarán en esta investigación. El objetivo de este análisis es el de obtener información significativa acerca de la problemática planteada. Posteriormente, se determinará la calidad de los datos bajo la suposición de que pueden presentar errores o incoherencias, lo que disminuiría su capacidad para cumplir los objetivos de la investigación.

2.2.1 Conjunto de datos inicial

La problemática que se plantea en esta investigación contempla la construcción de modelos para la predicción de la salud fetal y la posterior implementación de algoritmos de XAI. Para llevar a cabo una predicción razonablemente precisa y la extracción de características influyentes, es necesario disponer de un método de análisis de la salud fetal confiable y, de ser posible, debidamente probado. Como se indicó en el Capítulo 1, la cardiotocografía es una técnica utilizada por los profesionales de la salud para monitorizar la frecuencia cardíaca fetal y las contracciones uterinas de la madre durante las últimas

etapas de gestación [49]. Esta técnica puede ser aprovechada para obtener una idea general del estado de salud del no nacido.

Con la premisa mencionada anteriormente, se propone utilizar un conjunto de datos inicial que represente una descripción en formato tabular de cardiotocografías reales, esto permitirá crear modelos que simulen el análisis cardiotocográfico, pero haciendo uso de modelos inteligentes de aprendizaje automático y, a la vez, disponer de una serie de características diferenciadas entre sí, para establecer una sencilla interpretación de los datos.

El conjunto de datos inicial que se utilizará en esta investigación es el que se encuentra disponible en la base de datos del repositorio de UCI Machine Learning Repository [84] y consta de mediciones de las características de la frecuencia cardíaca fetal (FCF) y la contracción uterina (CU) en cardiotocografías clasificadas por expertos obstetras. El conjunto de datos ha sido propuesto por Ayres de Campos et al. [85] en año 2000 y donado al repositorio en 2010.

2.2.2 Descripción de los datos

A continuación, se presenta una descripción general del conjunto de datos:

- Formato: Tabular.
- Número de características: 22.
- Número de instancias: 2126.

En la Tabla 6 se provee una descripción de cada una de las características o atributos que conforman el conjunto de datos:

Tabla 6. Características del conjunto de datos y su descripción.

Característica	Descripción
baseline_value	Se expresa en latidos por minuto (lpm). Es la frecuencia cardíaca fetal (FCF) promedio redondeado a incrementos de 5 latidos por minuto durante un segmento de 10 minutos.
accelerations	Número de aceleraciones por segundo. Ascensos de la Frecuencia Cardíaca Fetal (FCF) de 15 a 25 latidos durante 15 segundos o más en relación con la FCF base.
fetal_movement	Número de movimientos del feto por segundo. Representa los movimientos fetales obtenidos ya sea por detección automatizada o registrados por la madre, dependiendo de las capacidades del monitor.
uterine_contractions	Número de contracciones uterinas por segundo. Las contracciones uterinas se definen como períodos que duran entre 20 y 240 segundos, en los que se perciben un endurecimiento del abdomen como consecuencia de la actividad del músculo uterino.
light_decelerations	Número de desaceleraciones ligeras por segundo. Las desaceleraciones representan una disminución de la FCF por debajo de la línea base, de más de 15 lpm de amplitud y durante más de 15 segundos

Característica	Descripción
prolongued_ decelerations	Número de desaceleraciones prolongadas por segundo. Las desaceleraciones se clasifican en prolongadas si duran entre 120 y 300 segundos.
severe_decelerations	Número de desaceleraciones severas por segundo. Las desaceleraciones se clasifican en severas si superan los 300 segundos.
abnormal_short_ term_variability	Porcentaje de tiempo con variabilidad anormal a corto plazo. Del 0 al 100%. La variabilidad se define como fluctuaciones en la línea de base de la FCF de 2 ciclos por minuto o más, con amplitud irregular y frecuencia inconstante. Un punto con variabilidad a corto plazo anormal es la diferencia entre dos señales de FCF adyacentes menores a 1 lpm.
mean_value_of_ short_term_variability	Valor medio de la variabilidad a corto plazo.
percentage_of _time_with_abnormal _long_term_ variability	Porcentaje de tiempo con variabilidad anormal a largo plazo (LTV). Representa la oscilación de la FCF alrededor de la línea de base en una amplitud de 5 a 10 lpm. La variación a largo plazo solo se evalúa en los segmentos que no se consideraron aceleraciones o desaceleraciones.
mean_value_of_ long_term_variability	Valor medio de variabilidad a largo plazo.
histogram_width	Representa el tamaño total del histograma de la frecuencia cardíaca fetal, en el cual se agrupan todos los cuadros de tiempo para representar el histograma completo.
histogram_min	Muestra el valor mínimo de la frecuencia cardíaca fetal representada en el histograma.
histogram_max	Muestra el valor máximo de la frecuencia cardíaca fetal representada en el histograma.
histogram_number_ of_peaks	Representa el número de picos que tiene el histograma, los cuales corresponden a los valores con mayor frecuencia.
histogram_number_ of_zeroes	Representa el número de veces que los valores del histograma llegan a cero.
histogram_mode	Representa la moda, lo cual es el valor más frecuente a lo largo del histograma.
histogram_mean	Representa la suma de los valores de todos los datos de la frecuencia cardíaca fetal, dividida entre el número de datos de la frecuencia cardíaca fetal.
histogram_median	Representa el valor central que del histograma.
histogram_variance	Representa la varianza, es decir, la variabilidad que tienen los datos del histograma.
histogram_tendency	Este valor representa la tendencia del histograma.
fetal_health	Representa la salud del feto, donde, 1 significa que el feto se encuentra con una salud normal, 2 significa que el feto tiene una salud que indica sospecha de enfermedad o riesgo y, finalmente, 3 significa que el feto tiene una salud patológica, es decir, que constituye una enfermedad o indica síntomas de ella.

2.2.3 Exploración de datos

La visualización de los datos facilita su comprensión y brinda una visión más profunda del estado de las características que conforman el conjunto de datos inicial. Además, las operaciones estadísticas fundamentales pueden apoyar este proceso de inspección y exploración.

Como se indicó en la Tabla 6 de la descripción de los datos, las características permiten clasificar la salud del feto en 3 categorías específicas. La Figura 13 muestra la cantidad de casos de cada una de las categorías. Es evidente que la mayoría de los casos corresponden a instancias con salud fetal normal, casi cuatro veces más que los casos con salud fetal sospechosa y patológica. Esto puede ser un indicador de un desbalance en el conjunto de datos inicial.

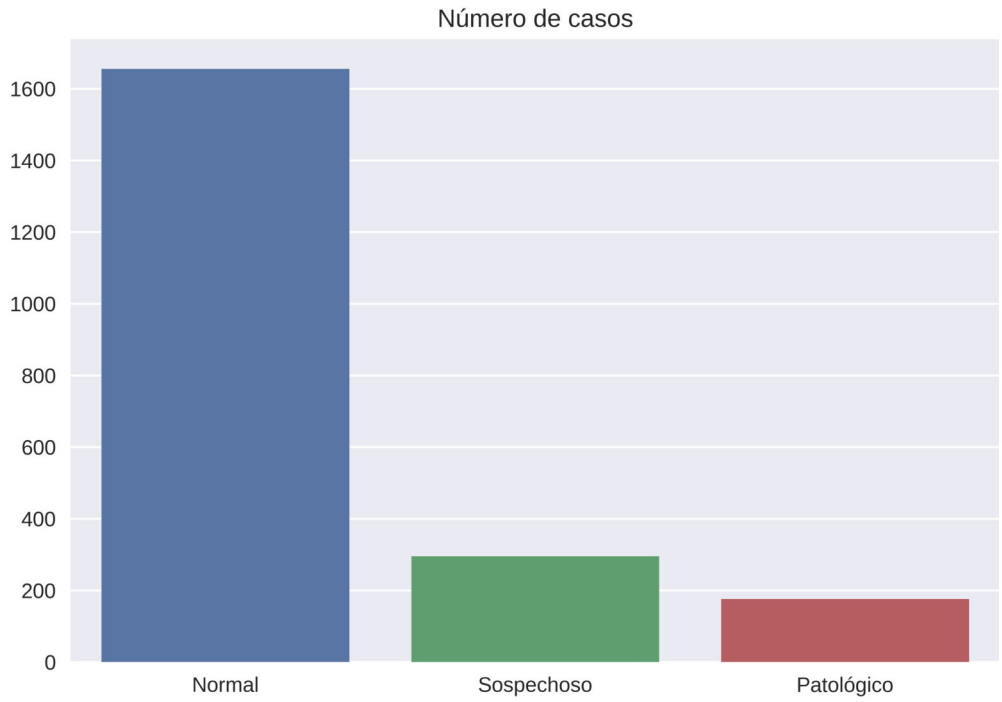


Figura 13. Número de casos por categoría de salud del feto.

En la Figura 14 se muestran los histogramas de los datos para representar la distribución de las características continuas que conforman el dataset inicial. Respecto a la distribución de los atributos por su forma, se puede mencionar que no existe una tendencia general. El atributo correspondiente a la línea base de la frecuencia cardíaca fetal presenta una forma aproximadamente simétrica, lo que indica una distribución normal; por el contrario, atributos como las aceleraciones o las desaceleraciones ligeras presentan una distribución sesgada hacia la derecha. Adicionalmente, un análisis preliminar de la variabilidad entre los datos se puede llevar a cabo observando cada uno de los histogramas con detenimiento.

Con la finalidad de conocer el grado de relación lineal entre los atributos del conjunto de datos se graficó una matriz de correlación, esta se presenta en la Figura 15. Los coeficientes de correlación toman valores entre -1 y 1 para representar relaciones negativas y positivas, respectivamente. Cabe mencionar que las comparaciones entre las mismas variables resultan en una correlación perfecta, por lo que adoptan el coeficiente 1.

Para facilitar la lectura de la matriz, los valores se muestran en forma de un mapa de calor en donde es posible identificar el grado de correlación con los colores de cada celda.

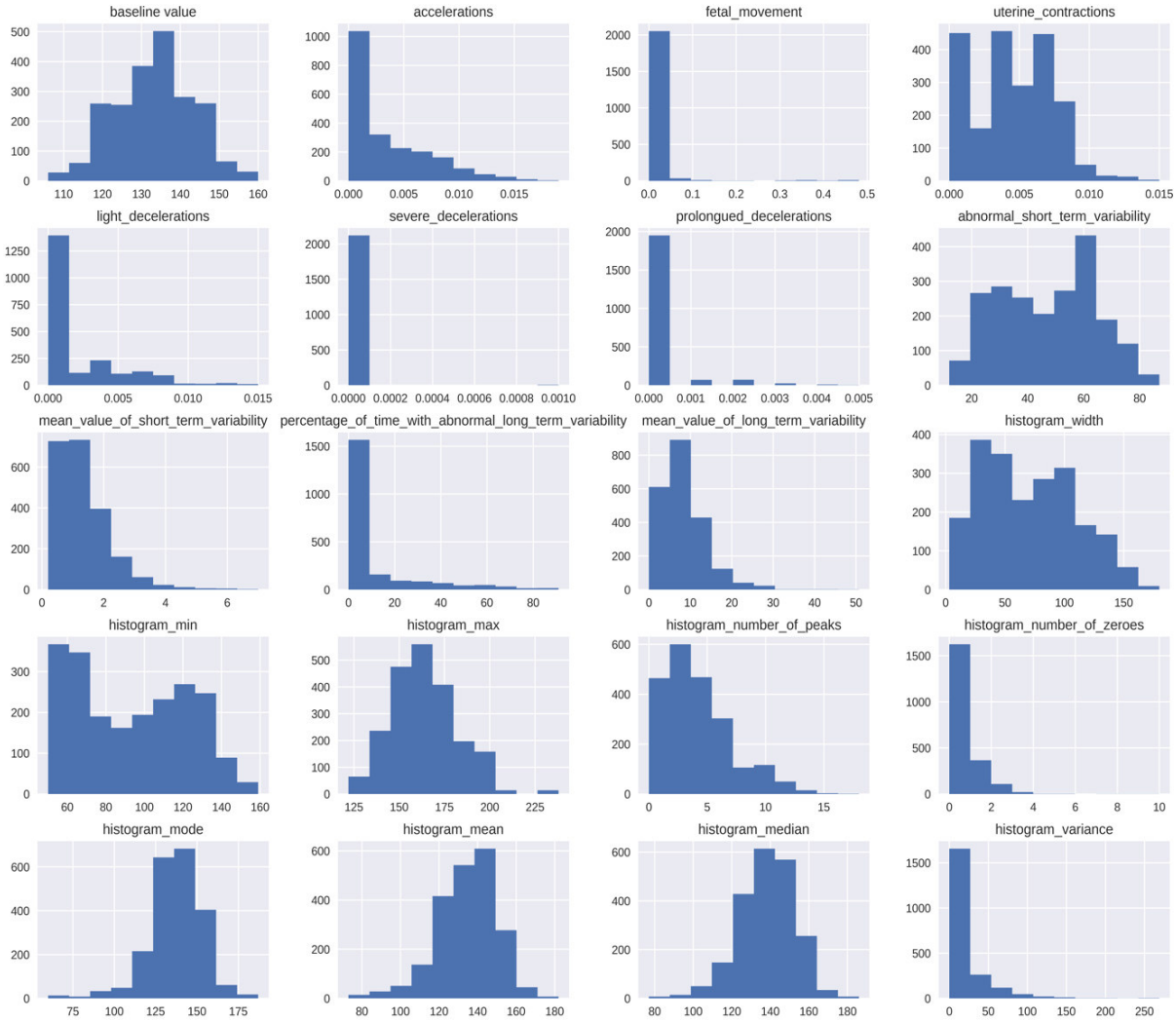


Figura 14. Histograma de los atributos del conjunto de datos.

Las relaciones que nos interesan primordialmente son aquellas que involucran a la variable *fetal_health*, pues esta dicta la clasificación de la salud fetal. Si tomamos los coeficientes de correlación de esta variable podemos notar que existe una correlación positiva con la variable *prolonged_decelerations* con un coeficiente de 0.48, lo que podría considerarse como una correlación moderada; otro caso es el de la variable *accelerations* que presenta una correlación negativa débil con un coeficiente de -0.36. De esta manera podemos interpretar la matriz de correlación y determinar la dependencia de los atributos de interés en el conjunto de datos de esta investigación.

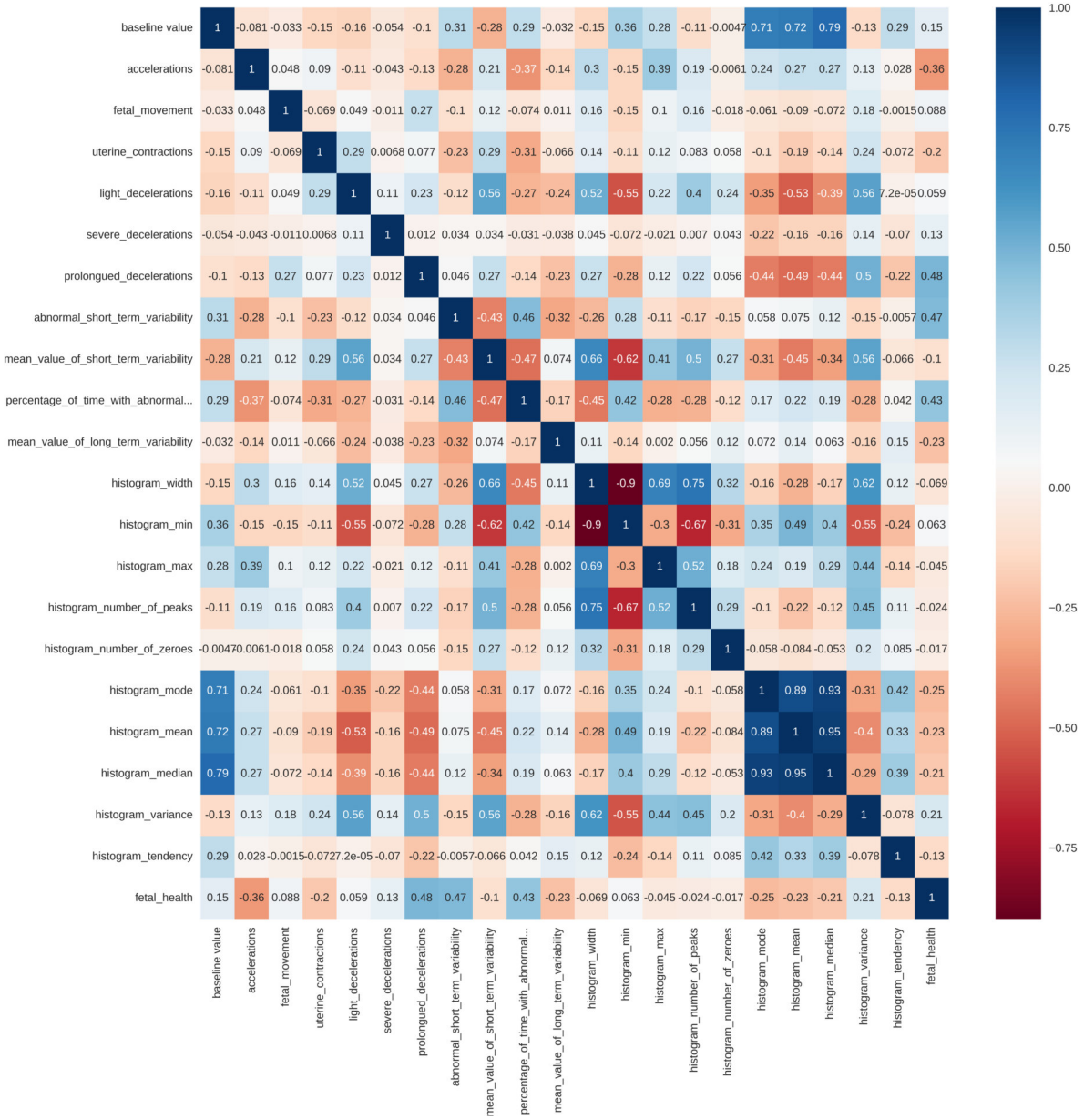


Figura 15. Matriz de correlación de las características del conjunto de datos.

2.2.4 Verificación de calidad de datos

Los conjuntos de datos suelen presentar imperfecciones que disminuyen su capacidad para cumplir los objetivos de la investigación. Estas imperfecciones pueden ser errores de congruencia en los datos que se cometieron al momento de la recolección, como un apartado del estado civil que presenta incoherencia en su escritura (el uso de S y Soltero para designar el mismo valor), otra imperfección bastante común corresponde a los errores tipográficos como las faltas ortográficas que podrían aparecer en los registros; sin embargo, estos errores se presentan únicamente cuando existen variables que no son

numéricas, este no es el caso para las variables de entrada del dataset utilizado; en la Figura 16 se puede apreciar que las variables son de tipo flotante.

```
baseline value                                                  float64
accelerations                                                   float64
fetal_movement                                                  float64
uterine_contractions                                            float64
light_decelerations                                             float64
severe_decelerations                                            float64
prolongued_decelerations                                        float64
abnormal_short_term_variability                                 float64
mean_value_of_short_term_variability                            float64
percentage_of_time_with_abnormal_long_term_variability          float64
mean_value_of_long_term_variability                             float64
histogram_width                                                 float64
histogram_min                                                   float64
histogram_max                                                   float64
histogram_number_of_peaks                                       float64
histogram_number_of_zeroes                                      float64
histogram_mode                                                  float64
histogram_mean                                                  float64
histogram_median                                                float64
histogram_variance                                              float64
histogram_tendency                                              float64
```

Figura 16. Tipo de dato de cada característica del conjunto de datos.

Un tipo de imperfección que se suele pasar por alto es el error de medición, en donde los valores son correctos, pero no se asume un esquema de medición homogéneo para todos los registros. En este caso, determinar estos errores no es una tarea sencilla, sin embargo, no se espera la presencia de este tipo de errores en el conjunto de datos utilizado en esta investigación, pues está respaldado por el proceso de recolección llevado a cabo por el autor del dataset [85] e incluso ha sido clasificado por los obstetras expertos.

Con los factores expuestos anteriormente se considera que el conjunto de datos es capaz de servir como entrada para cumplir los objetivos de la investigación.

2.3 Preparación de los datos

Para hacer uso de los datos obtenidos, estos pasarán por un proceso de preparación previa que involucra una limpieza en caso de existir datos erróneos o faltantes, el formateo de los datos según la necesidad de los modelos a implementar y una selección de características en función del aporte a la clasificación.

2.3.1 Limpieza de datos

En la etapa anterior se verificó la calidad del conjunto de datos utilizado en la investigación, sin embargo, aún existen algunos aspectos relevantes que se deben tomar en cuenta respecto a la presencia de posibles imperfecciones a ser solventadas antes de empezar

con la construcción de modelos. Principalmente, estos desperfectos se relacionan con la existencia de valores nulos o la misma ausencia de valores, es decir, registros vacíos.

Tabla 7. Número de valores nulos y vacíos de cada característica del conjunto de datos.

Característica	Nulos	Vacíos
baseline value	0	0
accelerations	0	0
fetal_movement	0	0
uterine_contractions	0	0
light_decelerations	0	0
severe_decelerations	0	0
prolongued_decelerations	0	0
abnormal_short_term_variability	0	0
mean_value_of_short_term_variability	0	0
percentage_of_time_with_abnormal_long_term_variability	0	0
mean_value_of_long_term_variability	0	0
histogram_width	0	0
histogram_min	0	0
histogram_max	0	0
histogram_number_of_peaks	0	0
histogram_number_of_zeroes	0	0
histogram_mode	0	0
histogram_mean	0	0
histogram_median	0	0
histogram_variance	0	0
histogram_tendency	0	0
fetal_health	0	0

Para indicar que el conjunto de datos no presenta el tipo de imperfecciones descrito anteriormente, la Tabla 7 muestra el número de registros que presentan valores nulos y vacíos.

2.3.2 Formato de datos

Es posible que algunas de las técnicas de aprendizaje automático que se utilizarán posteriormente requieran de un formato específico para su implementación. Generalmente, estos formatos pueden representar un orden específico o una clasificación determinada de los datos. Debido a que el dataset utilizado en esta investigación se utilizará para la clasificación de la salud fetal y, de hecho, existe un atributo que indica esta característica en cada registro, los modelos a implementar serán del tipo supervisado.

Con este antecedente, el conjunto de datos será dividido en dos subconjuntos para su manipulación. El primer subconjunto, que será llamado *X*, corresponde a los parámetros de entrada para la clasificación; mientras que el segundo subconjunto, *Y* en este caso, incluye el atributo *fetal_health* que representa el valor de la salud fetal de los registros, es decir, la salida de la clasificación. Esta separación, además de servir como entrada para los modelos, permitirá efectuar otras técnicas de preparación de datos como las que se describen posteriormente.

2.3.3 Selección de características

El conjunto de datos inicial suele presentar atributos que son irrelevantes para los objetivos del proyecto. En muchos de los casos es necesario seleccionar únicamente aquellas características que influyen en el resultado, esto permite ignorar los atributos que actúan como ruido en el modelo de aprendizaje automático y mejorar su rendimiento predictivo. Inclusive, la selección de características podría aumentar la capacidad explicativa de los modelos de XAI al disponer de menos factores involucrados en las decisiones a interpretar.

Para solventar esta problemática, se implementó una técnica conocida como selección de características univariadas que funciona mediante la elección de las mejores características basándose en pruebas estadísticas. Estas pruebas tienen la particularidad de involucrar una única variable dependiente, de lo contrario se convierte en análisis bivariado en el caso de analizar dos variables o multivariado en el caso de analizar más. En la práctica, se utilizó la función SelectKBest de la librería Scikit-learn [86] para seleccionar una serie de características que conservar.

La función SelectKBest recibe como parámetro el test que evaluará las características y devolverá puntuaciones univariadas. En este caso, el test utilizado corresponde al denominado *f_classif*, debido principalmente a que la variable objetivo es una clasificación de la salud fetal en los registros; para casos donde la salida es una regresión lineal existe el test *f_regression*. Estos métodos estiman el grado de dependencia lineal entre dos variables aleatorias y le asignan un puntaje específico a cada una.

Por cuestiones de interpretabilidad, las características del dataset inicial que serán puestas a prueba para determinar aquellas que pueden ser descartadas corresponden únicamente a las que describen elementos de los histogramas de las cardiotocografías. Esto evitará que se pierda información relevante respecto a los factores involucrados en la frecuencia cardíaca fetal y las contracciones uterinas. El resultado se indica en la Figura 17.

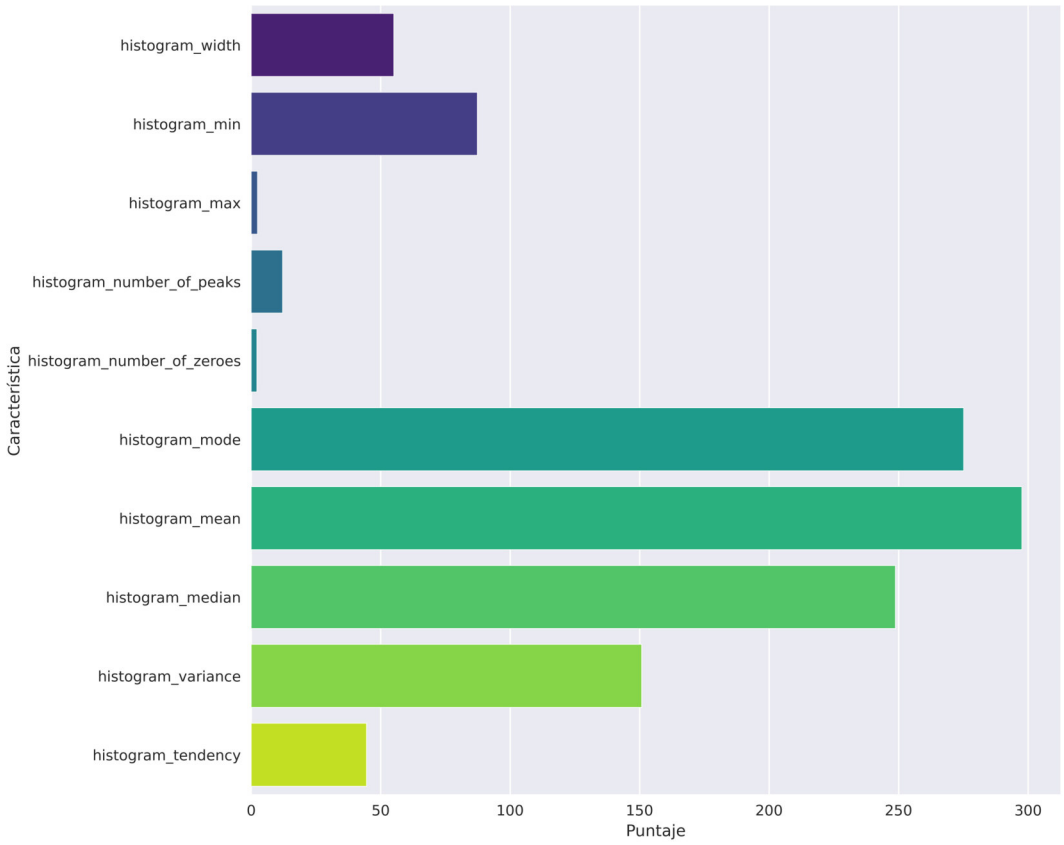


Figura 17. Resultado de la selección de características mediante la función SelectKBest.

El puntaje de las características evaluadas no es homogéneo, podemos observar variables como el promedio del histograma que presentan una fuerte dependencia lineal con la variable objetivo. Por el contrario, variables como el número de ceros en el histograma o el número de picos no parecen aportar un valor real al resultado, incluso podrían causar ruido en la construcción de modelos. Por esta razón, se decidió incluir en el conjunto de datos refinado únicamente a las características que hayan obtenido un puntaje mayor a 100 en el resultado de la función SelectKBest, estas características son:

- histogram_mode
- histogram_mean
- histogram_median
- histogram_variance

2.3.4 Estandarización de características

Existe un último paso necesario para evitar errores en la construcción de modelos. Cuando una característica presenta una varianza muy distinta a las demás, esta obstaculiza la comparación directa entre variables, lo que podría afectar el aprendizaje del modelo. Para

evitar este problema es necesario aplicar una estandarización a los datos que suele presentarse como un escalado de la varianza. En la práctica se utilizó la función StandarScaler de la librería Scikit-learn [86], que estandariza las características restando la media y escalando a la unidad la varianza de cada una de las muestras que recibe.

En la Figura 18 se muestra la distribución de las variables de entrada mediante un diagrama de cajas, se puede notar la diferencia en la dispersión de datos entre las variables. La Figura 19 muestra la distribución luego de aplicar la técnica de StandarScaler. Es evidente la diferencia en los resultados, la uniformidad de los datos se agudiza y se eliminan las desviaciones mayores que obstaculizan el aprendizaje de los modelos. Cabe destacar que la estandarización de datos se aplicará únicamente para los algoritmos de AI que lo requieran.

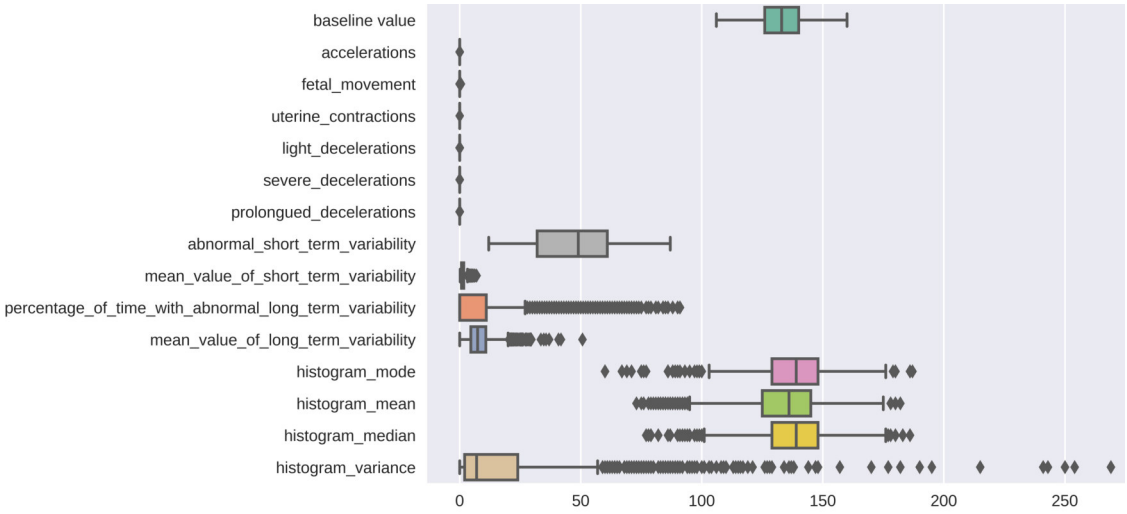


Figura 18. Diagrama de caja de la distribución de las variables del conjunto de datos.

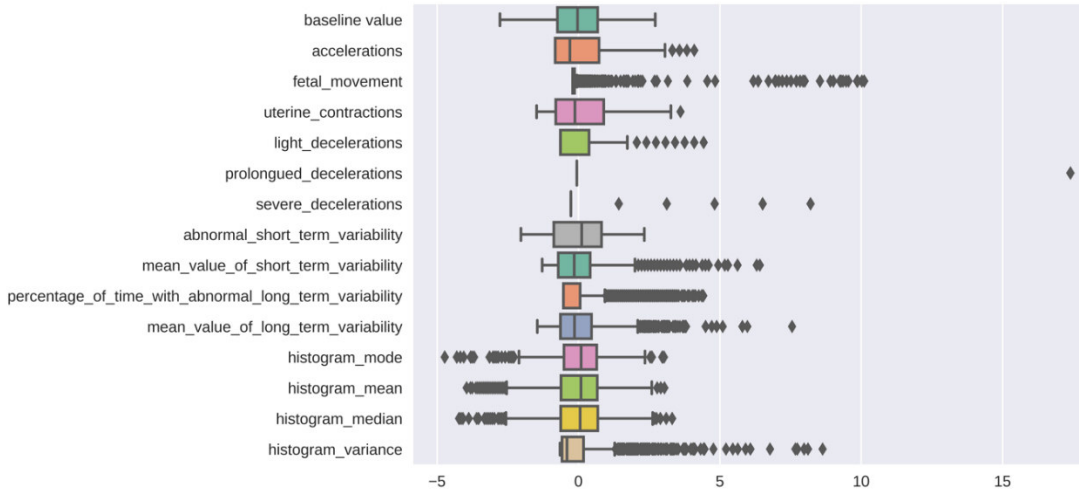


Figura 19. Diagrama de caja de la distribución de las variables del conjunto de datos después de aplicar la técnica de estandarización.

2.4 Modelado

Esta fase no consiste únicamente en la construcción de modelos, en realidad, comprende una serie de pasos que permiten determinar la factibilidad de los algoritmos para alcanzar los objetivos planteados. En primer lugar, se llevará a cabo un proceso de selección para determinar los algoritmos de AI que permitirán predecir la salud fetal con los datos de entrada. A continuación, se diseñará el plan de prueba para identificar la precisión de los algoritmos de aprendizaje automático mediante métricas de rendimiento. Adicionalmente, se considerará una separación de datos en conjuntos de entrenamiento, prueba y validación.

Los modelos serán construidos con los parámetros requeridos por cada algoritmo y sus salidas serán evaluadas según el plan de prueba definido. Una vez obtenidos los resultados de los modelos de aprendizaje automático, se construirá el modelo de XAI para la interpretación de las decisiones tomadas por el algoritmo de AI que haya dado los mejores resultados. La técnica de XAI que se implementará es LIME y sus salidas se evaluarán en la siguiente fase de la metodología, considerando los objetivos del negocio.

2.4.1 Selección de técnicas de AI

Un problema de análisis de datos puede ser abordado con distintas técnicas, sin embargo, es necesario tomar en cuenta ciertos factores para determinar aquellos algoritmos de aprendizaje automático que se adecuan favorablemente al contexto. En primer lugar, la configuración de los datos consiste en características de entrada y salida, por esta razón, se estableció que los algoritmos a implementar aprenderán de manera supervisada.

Otro factor que se debe tomar en cuenta en este caso es la naturaleza del algoritmo de XAI que se implementará posteriormente. Dado que LIME es una técnica que se aplica a los resultados, es decir, es de naturaleza “post-hoc”, los algoritmos de aprendizaje automático que se beneficiarán de su capacidad explicativa son los de tipo “black box”. Por esta razón se descartan los algoritmos intrínsecamente explicables como regresiones lineales o logísticas.

Con estos antecedentes, los algoritmos de aprendizaje automático que se implementarán en este proyecto son los siguientes:

- Máquina de vectores de soporte (SVM, por sus siglas en inglés).
- Random Forest.
- Redes Neuronales Artificiales (ANN, por sus siglas en inglés).

2.4.2 Diseño de plan de pruebas

Si deseamos saber cuál es el modelo de aprendizaje automático que otorga los mejores resultados, debemos disponer de una serie de pruebas que comprueben objetivamente su rendimiento. Las pruebas se ejecutarán en iguales condiciones para cada uno de los modelos con la finalidad de establecer comparaciones objetivas y determinar el mejor algoritmo para el problema planteado.

En primer lugar, es necesario tomar en cuenta que el conjunto de datos no se encuentra balanceado y las métricas que se utilizarán deben soportar esta particularidad. Con este antecedente es posible descartar el uso de la métrica “accuracy”, que puede presentar inexactitudes al momento de precisar el rendimiento de un modelo entrenado con datos desbalanceados. Para solventar este inconveniente, se propone utilizar matrices de confusión y la métrica F1-Score, que resultan más apropiadas para este tipo de situaciones.

El uso de matrices de confusión permitirá conocer de manera práctica los resultados de los algoritmos, pues esta herramienta permite visualizar el número de resultados correctos y erróneos que se emiten. Con estos datos es posible interpretar la verdadera capacidad de predicción del modelo, determinando la gravedad de sus errores representados mediante falsos positivos y falsos negativos. La métrica F1-Score será una herramienta que permitirá comparar los modelos basándose en los resultados de la matriz de confusión a través del promedio de los valores conocidos como Recall y Precision.

2.4.3 Construcción de modelos de aprendizaje automático

Máquina de vectores de soporte

El funcionamiento de esta técnica empieza por los elementos que le dan nombre, los llamados “vectores de soporte” [87]. Los datos etiquetados se separan en clases utilizando un límite de separación óptimo que se denomina hiperplano, pues divide un plano en dos secciones con una clase en cada una. Los vectores de soporte son los datos de clase más cercanos al hiperplano que sirven de guía para la optimización de la separación de clases mediante un “margen máximo”. Cuando el problema no es lineal, SVM transforma el espacio de representación de los datos de entrada a una dimensión mayor mediante funciones “kernel”, donde los datos se pueden separar con un hiperplano lineal.

La implementación de este algoritmo de aprendizaje automático se llevó a cabo mediante la librería SVM de Scikit Learn [86]. Específicamente, se utilizó la clase “SVC” (Support Vector Classification) cuya funcionalidad consiste en la clasificación binaria o de múltiples clases en un determinado conjunto de datos. Se ocuparon los parámetros por defecto, a

excepción del denominado “random_state”, que recibe un entero para garantizar la persistencia en la mezcla de datos, esto permitirá obtener resultados reproducibles en cada ejecución.

Random Forest

El algoritmo de bosques aleatorios consiste, a grandes rasgos, en el ensamblaje de árboles de decisión independientes, lo que genera un bosque de decisiones que da el nombre a esta técnica [88]. La naturaleza aleatoria del algoritmo viene dada por un sorteo de las instancias y las características de los datos que servirán de entrada para cada árbol, en donde cada uno emitirá una decisión independiente. De esta manera, la predicción hecha por el bosque aleatorio para datos desconocidos corresponde a la clase más votada por todos los árboles.

La clase RandomForestClassifier de Scikit Learn [86] permitió implementar este algoritmo en el proyecto. Se utilizaron los valores por defecto en los parámetros de la clase, por ejemplo, el parámetro “bootstrap” presenta el valor predeterminado “true” y permite utilizar submuestras del conjunto de datos en la construcción de árboles para controlar el sobreajuste, disminuyendo riesgos que puedan afectar la precisión del modelo. En este caso en concreto, no fue necesario aplicar la estandarización de datos, puesto que los algoritmos basados en árboles se implementan a partir de reglas de decisión en lugar de un cálculo de distancias [89].

Red Neuronal Artificial

Las redes neuronales artificiales buscan emular el funcionamiento del cerebro humano con diferentes capas de neuronas que clasifican y etiquetan las características de los datos de entrada para formular el conocimiento que permite ejecutar predicciones [90]. La estructura de una red neuronal empieza por la primera capa que recibe los datos y realiza un cálculo con una función de activación específica para producir un resultado, generalmente una predicción probabilística. Este resultado se transmite a la siguiente capa de neuronas con un peso asociado que define la influencia de la neurona en las neuronas adyacentes y posiblemente en el resultado final.

En este caso, la implementación de esta técnica de caja negra se llevó a cabo con la clase “Dense” de la librería “Keras” de TensorFlow [91]. Las capas son densas debido a la posible existencia de relaciones entre cualquier atributo del conjunto de datos, lo que permite que los nodos estén totalmente conectados. La arquitectura de la red neuronal implementada se indica en la Figura 20.

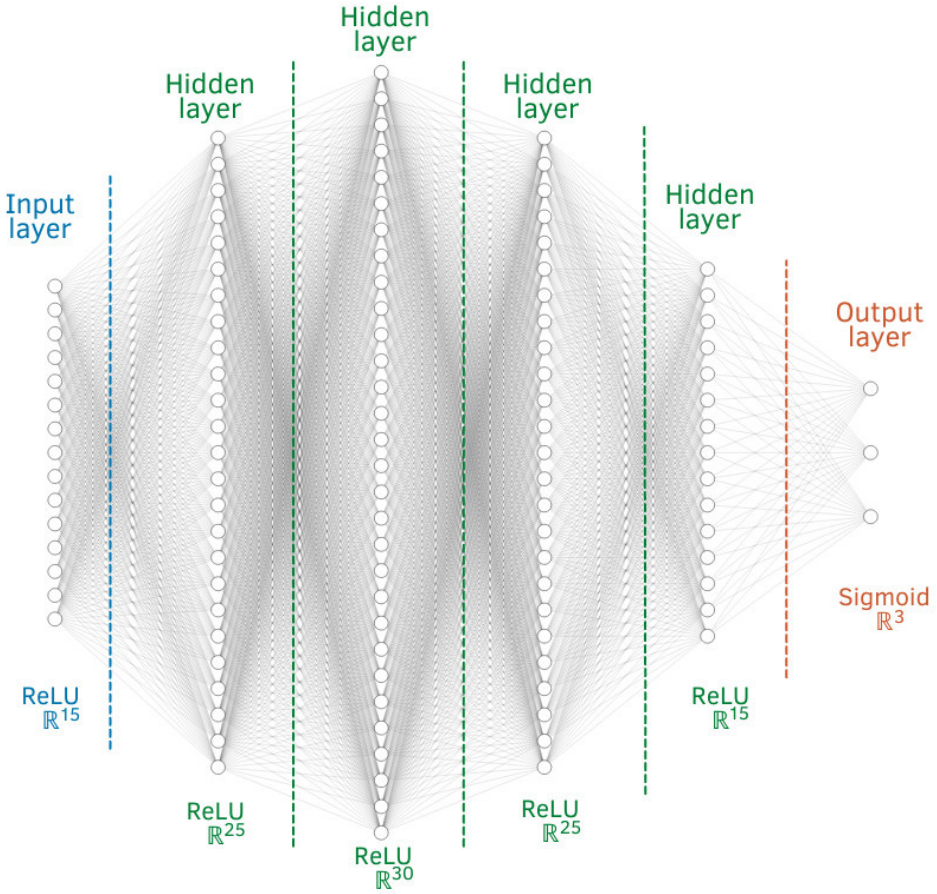


Figura 20. Arquitectura de la red neuronal densamente conectada.

Las capas de entrada e intermedias (hidden) utilizan la función de activación lineal rectificada (ReLU) para que los datos transformados por la función sean linealmente separables y faciliten la construcción del modelo de clasificación. La función de activación "softmax" se utiliza en la capa "output" por su capacidad para emitir las salidas del modelo de clasificación de múltiples clases mutuamente excluyentes, pues un feto no puede tener varios estados (normal, patológico o sospechoso) a la vez; esta función no se utiliza en capas intermedias para evitar errores como el desvanecimiento del gradiente [92].

2.4.4 Evaluación de modelos

Es momento de evaluar los algoritmos de aprendizaje automático implementados y seleccionar aquel que presente los mejores resultados. Con esta finalidad, se utiliza el plan de pruebas diseñado anteriormente. En esta sección se utilizarán las métricas de Precision, Recall y F1-Score para plantear una comparación del rendimiento entre cada uno de los algoritmos. Más detalles se pueden observar en el Capítulo 3, donde se especifican los criterios objetivos y subjetivos que determinan el mejor modelo.

Tabla 8. Resultados del rendimiento de los algoritmos de aprendizaje automático para las clases Normal (N), Sospechoso (S) y Patológico (P).

	Precision			**Recall**			**F1-Score**		
	N	S	P	N	S	P	N	S	P
SVM	0,95	0,76	0,87	0,95	0,74	0,83	0,95	0,75	0,85
RF	0,96	0,91	0,93	0,98	0,79	0,93	0,97	0,85	0,93
ANN	0,94	0,79	0,72	0,95	0,67	0,88	0,95	0,73	0,79

La Tabla 8 muestra los resultados que obtuvieron los algoritmos con las métricas mencionadas anteriormente, esto para cada una de las posibles salidas de la clasificación. Se puede observar que el clasificador del modelo de Random Forest obtiene los valores más altos en todas las instancias. Así, este modelo se selecciona como aquel que otorga los mejores resultados en la clasificación de la salud fetal y sus decisiones serán interpretadas con el algoritmo de XAI que se implementará a continuación.

2.4.5 Implementación del algoritmo de XAI

La interpretación del modelo de aprendizaje automático con los mejores resultados, Random Forest en este caso, se llevará a cabo con el algoritmo de Inteligencia Artificial Explicada LIME. La implementación se realiza con el módulo “lime_tabular” [19] que contiene las funciones de explicación para aquellos clasificadores que utilizan datos tabulares como el conjunto de datos de este proyecto.

Dentro de este módulo se encuentra la clase “LimeTabularExplainer” que construye el objeto explicador mediante la configuración de ciertos parámetros. Para este caso, la configuración es la siguiente:

Tabla 9. Parámetros de la clase "LimeTabularExplainer" para la construcción del explicador.

Parámetro	**Valor**	**Descripción**
training_data	np.array(X_train)	La función recibe como parámetro los datos de entrenamiento en forma de array.
feature_names	X_train.columns	Se registran los nombres de las columnas de los datos de entrenamiento.
class_names	['Normal','Sospechoso', 'Patológico']	Se asigna un nombre a las clases de salida.
verbose	True	Imprime los valores de predicción locales del modelo de ajuste lineal.
mode	"classification"	Se indica que es un caso de clasificación (no de regresión).

Esta representa la configuración inicial para el objeto que servirá de base en la ejecución de explicaciones locales. Sin embargo, es posible notar que no se ha determinado la

instancia específica que se desea interpretar, ni el algoritmo de aprendizaje automático del cual se evaluarán las decisiones obtenidas.

Luego de seleccionar una instancia al azar del conjunto de datos de prueba y colocarla en formato de arreglo, utilizamos la función “explain_instance” del objeto creado anteriormente, esto permite generar las explicaciones de la clasificación para la instancia determinada. Los parámetros utilizados en esta función se describen a continuación:

Tabla 10. Parámetros de la función "explain_instance" para la generación de explicaciones de una instancia determinada.

Parámetro	Valor	Descripción
data_row	instancia_determinada	Instancia de entrada que será interpretada.
predict_fn	random_forest.predict_proba	Función de predicción del modelo de aprendizaje automático cuyas decisiones serán evaluadas, Random Forest en este caso.
num_features	10	Número de características más influyentes que aparecerán en la explicación.
labels	[0, 1, 2]	Clases de salida que aparecerán en la explicación.

Los resultados de la ejecución de la función descrita anteriormente se detallan en el Capítulo 3.

2.5 Evaluación del modelo de XAI

Para evaluar la capacidad explicativa del modelo de XAI, LIME en este caso, es necesario aplicar un enfoque más subjetivo que las métricas de precisión utilizadas con los algoritmos de aprendizaje automático. Las salidas de modelo de XAI se presentan en forma de explicaciones que deben ser comprendidas por el usuario final, por lo tanto, su intervención es fundamental para determinar la validez del algoritmo.

Bajo esta premisa, se ha decidido utilizar el enfoque propuesto en el Modelo de Aceptación de Tecnología (TAM, por sus siglas en inglés) [93], que facilita la evaluación de tecnología y sistemas de información desde un punto de vista subjetivo. TAM basa sus criterios de aceptación en dos principales premisas:

- Facilidad de uso percibida: percepción del usuario referente al nivel de dificultad que se le presenta cuando hace uso del sistema; si la dificultad es alta, el beneficio que recibe el usuario con el sistema se anula por la imposibilidad de utilizarlo.
- Utilidad percibida: percepción del usuario referente al grado en el que los beneficios del sistema facilitan su labor; si el sistema le resulta indiferente al usuario, sus beneficios no son relevantes.

La implementación de este modelo de aceptación se llevó a cabo mediante una entrevista semiestructurada a un usuario determinado. Se definió un conjunto de preguntas de carácter abierto con la finalidad de responder a las premisas descritas anteriormente. Las preguntas se listan a continuación:

Facilidad de uso percibida:

1) ¿Qué tan fácil de entender le parecieron las explicaciones obtenidas?
2) ¿Con qué facilidad podría implementar estas explicaciones en su trabajo?

Utilidad percibida:

1) ¿Qué tan útiles le parecieron las explicaciones obtenidas?
2) ¿Qué tan útil considera que sería implementar en su trabajo las explicaciones obtenidas?

Adicionalmente, se propuso una pregunta para conocer los factores que podrían mejorar los resultados del algoritmo según el criterio del entrevistado:

- ¿Qué recomendaciones daría para mejorar las explicaciones obtenidas?

El usuario seleccionado para ser entrevistado cumple con los criterios que definen al público objetivo de este sistema, es decir, profesionales de la salud especializados en el área de obstetricia que traten asuntos relacionados con la salud fetal. En términos más específicos, el perfil del usuario entrevistado se describe a continuación:

> Médica general, con especialidad en Ginecología y Obstetricia. Formación en Patología del tracto genital inferior. Maestría en Climaterio y Menopausia. Autora de Caso Clínico.

El proceso de entrevista inicia con una demostración práctica de los modelos de aprendizaje automático y XAI, indicando gráficamente los resultados de la clasificación y las explicaciones otorgadas, esto con una instancia para cada clase de la variable de salida. A continuación, se plantean al usuario las preguntas de evaluación con enfoque TAM de manera secuencial, guardando un registro de las respuestas. Los resultados de la evaluación se detallan en el Capítulo 3.

2.6 Despliegue

Luego de corroborar la validez de los algoritmos de aprendizaje automático y XAI, se diseñó un plan de despliegue de los resultados. Para esto, se presentan estrategias de

implementación y control que permitan establecer una ruta de procesos desde el recibimiento de datos hasta el monitoreo periódico del funcionamiento de los modelos.

Se plantea un proceso no lineal con actividades que faciliten el despliegue de los resultados de los modelos y, a su vez, procuren su mejora continua. Este proceso se puede observar en la Figura 21:

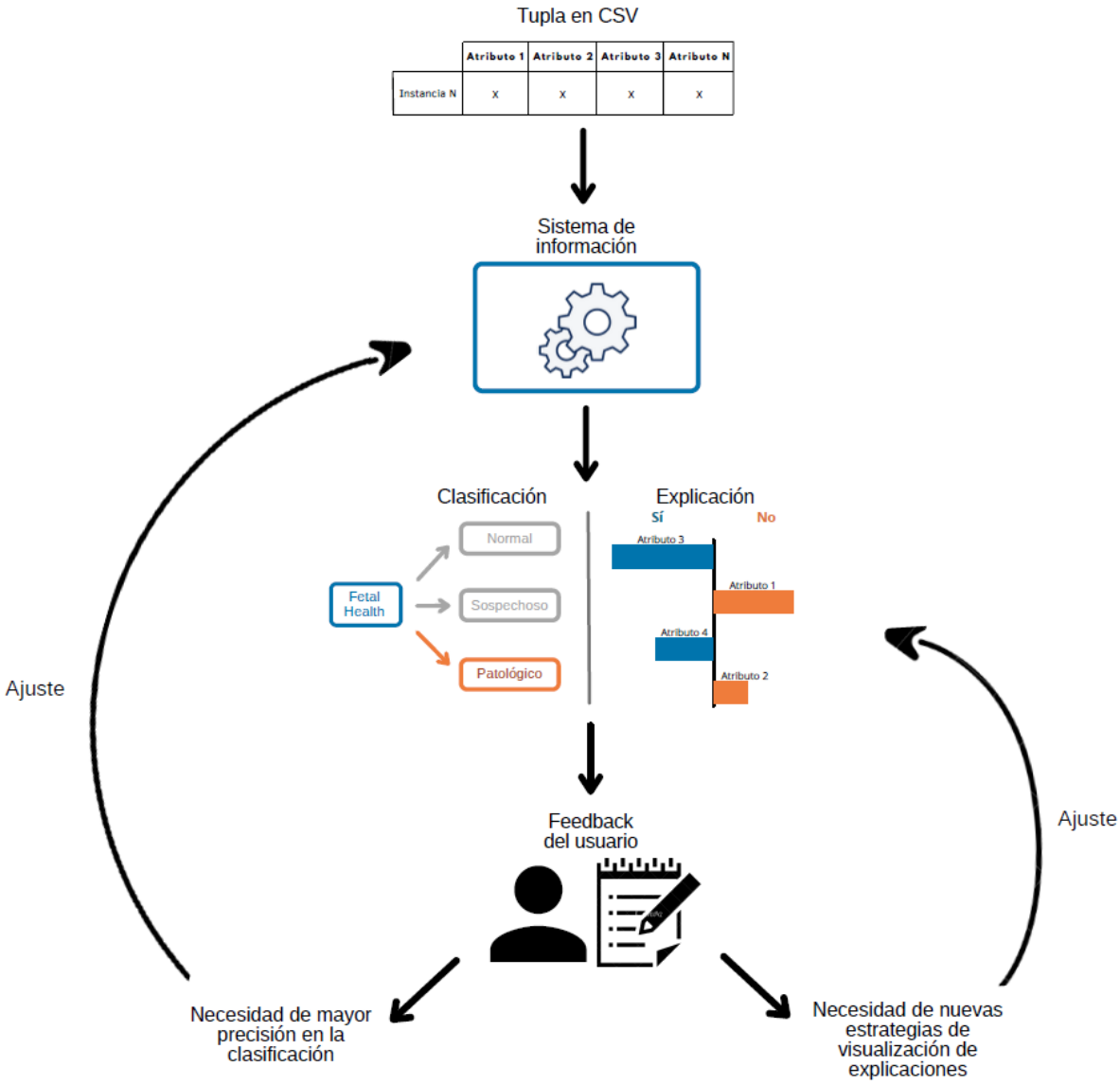


Figura 21. Estrategia de despliegue planteada para los algoritmos de AI y XAI en la toma de decisiones médicas para la salud fetal.

En primer lugar, es necesario definir el formato de los valores de entrada que serán evaluados por los modelos construidos en las etapas anteriores. Como se trata de un sistema de información que analiza datos en formato tabular, la entrada se constituye como fila que contiene los valores de cada variable en forma de columna, lo que representa una

instancia de datos. Para facilitar el registro de la entrada, el usuario puede recibir un formulario donde ingrese el valor de cada variable y, de manera interna, se transformen las respuestas del formulario en una tupla de datos en formato CSV.

A continuación, el sistema de información lee el archivo CSV que contiene la instancia introducida por el usuario. Esta es toda la información que requiere el sistema para funcionar. En primera instancia, el algoritmo de aprendizaje automático seleccionado (Random Forest) clasifica la instancia en una de las 3 clases de salida (“Normal”, “Sospechoso” o “Patológico”). Ahora, LIME permitirá interpretar las decisiones tomadas por el modelo de Random Forest otorgando al usuario explicaciones de manera gráfica. Se pueden agregar otras formas de visualización de las explicaciones si así se lo decide.

Cuando se ha hecho uso del sistema de información, es importante monitorizar su desempeño mediante la retroalimentación de la experiencia obtenida por el usuario. Este feedback puede ser recabado a través de un formulario de preguntas enfocadas en obtener información acerca de dos factores específicos:

Necesidad de mayor precisión en la clasificación: conocer la opinión del usuario acerca del rendimiento en la clasificación que otorga el sistema. El usuario puede considerar la existencia de un sesgo o inexactitud en este aspecto; dado el caso, se requerirá un ajuste en el algoritmo de aprendizaje automático.

Necesidad de nuevas estrategias de visualización de explicaciones: conocer la opinión del usuario acerca de la capacidad para interpretar las explicaciones mostradas por LIME. El usuario puede considerar la existencia de una alta complejidad en la interpretación de los resultados. Dado el caso, se requerirá un ajuste en la estrategia de visualización de explicaciones de LIME.

3 RESULTADOS Y CONCLUSIONES

3.1 Resultados

En esta sección se describen los resultados obtenidos con la implementación de la metodología. Específicamente, se detallan los resultados de la ejecución del plan de pruebas en los algoritmos de aprendizaje automático y las salidas que otorga el algoritmo de XAI en forma de explicaciones interpretables.

3.1.1 Resultados de la evaluación de modelos de aprendizaje automático

En el plan de pruebas diseñado en el capítulo anterior se plantearon las métricas que permitirán determinar aquel algoritmo que otorga los mejores resultados para este problema de clasificación. Para cada algoritmo de aprendizaje automático se muestran los resultados de la matriz de confusión (en forma porcentual para cada clase) y las métricas Precision, Recall y F1-Score.

Support Vector Machine

La matriz de confusión de los resultados otorgados por el algoritmo Support Vector Machine se presenta a continuación:

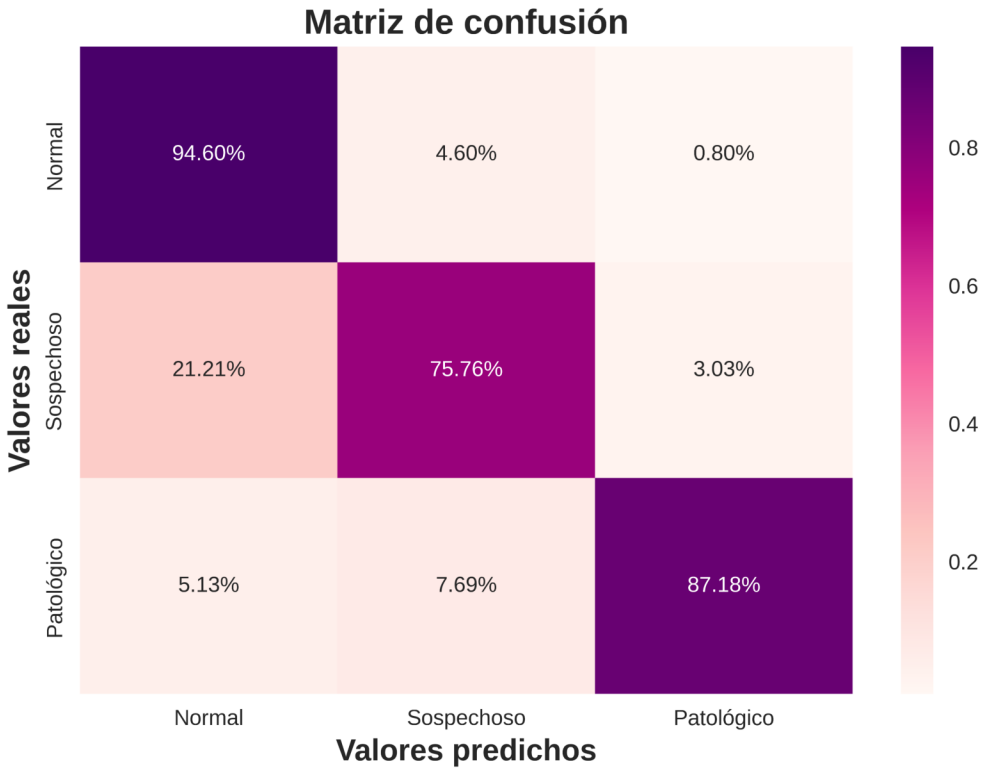


Figura 22. Matriz de confusión de los resultados del modelo de clasificación Support Vector Machine.

Las métricas de Precision, Recall y F1-Score del modelo SVM se visualizan en la Tabla 11:

Tabla 11. Resultado de las métricas Precision, Recall y F1-Score para el modelo de clasificación Support Vector Machine.

Support Vector Machine			
	Precision	**Recall**	**F1-Score**
Normal	0,95	0,95	0,95
Sospechoso	0,76	0,74	0,75
Patológico	0,87	0,83	0,85

Random Forest

La matriz de confusión de los resultados otorgados por el algoritmo Random Forest se presenta a continuación:

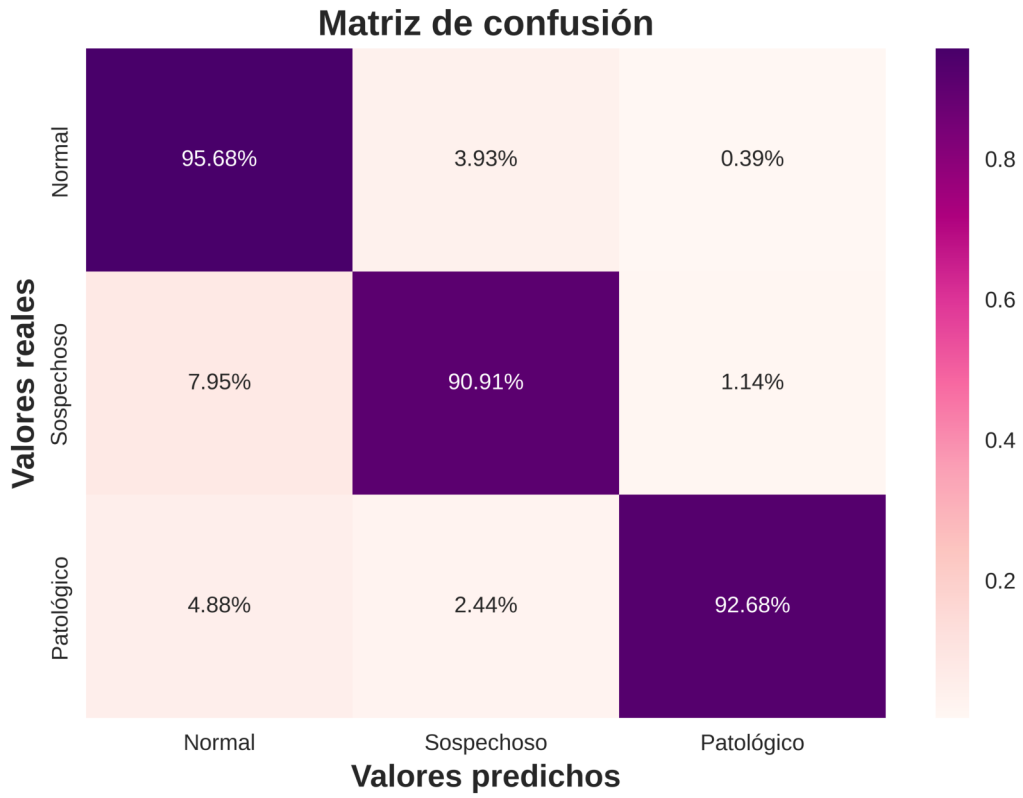


Figura 23. Matriz de confusión de los resultados del modelo de clasificación Random Forest.

Las métricas de Precision, Recall y F1-Score del modelo Random Forest se visualizan en la Tabla 12:

Tabla 12. Resultado de las métricas Precision, Recall y F1-Score para el modelo de clasificación Random Forest.

Random Forest			
	Precision	**Recall**	**F1-Score**
Normal	0,96	0,98	0,97
Sospechoso	0,91	0,79	0,85
Patológico	0,93	0,93	0,93

Red Neuronal Artificial

La matriz de confusión de los resultados otorgados por la red neuronal artificial se presenta a continuación:

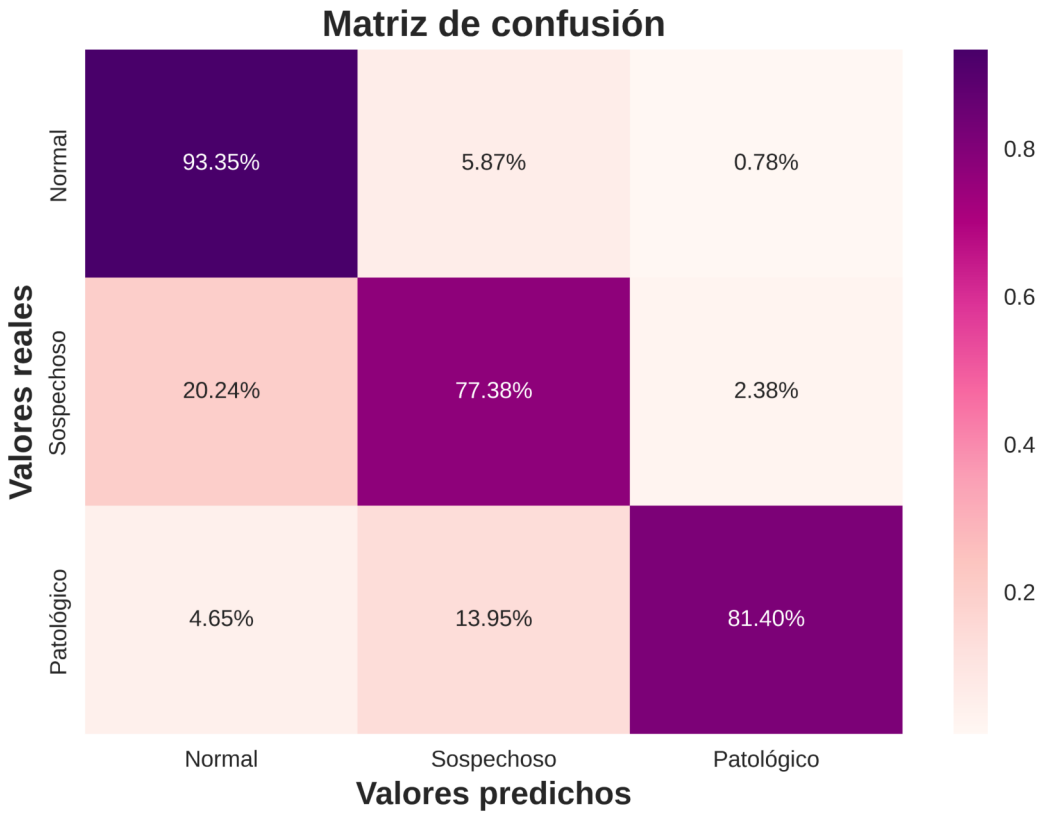


Figura 24. Matriz de confusión de los resultados del modelo de clasificación con ANN.

Las métricas de Precision, Recall y F1-Score del modelo de red neuronal artificial se visualizan en la Tabla 13:

Tabla 13. Resultado de las métricas Precision, Recall y F1-Score para el modelo de clasificación de Red Neuronal Artificial.

Red Neuronal Artificial			
	Precision	**Recall**	**F1-Score**
Normal	0,93	0,96	0,95
Sospechoso	0,77	0,64	0,70
Patológico	0,81	0,85	0,83

Las matrices de confusión muestran los resultados de manera porcentual respecto al número total de valores que pertenecen a una misma clase, para mejorar la visualización de la matriz se presenta una escala de color que crece en intensidad según el valor de la celda.

A simple vista, se puede notar que los tres algoritmos ofrecen resultados aceptables respecto a los valores estimados correctamente (diagonal de la matriz), pues ninguno de ellos disminuye del 60% del total. Sin embargo, la diagonal no es suficiente para determinar cuál es el mejor algoritmo, es necesario contextualizar los resultados al problema planteado.

Ya que el objetivo de estos algoritmos de aprendizaje automático consiste en la clasificación de la salud de un feto en tres posibles categorías, cada resultado conlleva una relevancia específica respecto al error, más aún si tomamos en cuenta la naturaleza de la solución como un soporte en el diagnóstico médico y toma de decisiones.

En términos prácticos, no es conveniente tolerar una situación en la que un paciente sea diagnosticado como “normal”, cuando en realidad presenta características problemáticas para su salud (“sospechoso” o “patológico”), pues retrasaría el acceso a un tratamiento adecuado. Por otro lado, que un paciente sea diagnosticado como “sospechoso” o “patológico”, cuando no tiene problemas de salud, representa un riesgo menor al anterior, pues no requiere de ningún tratamiento.

De esta manera, podemos concluir que lo que nos interesa es evitar los falsos negativos en la clasificación de registros. Por lo tanto, la métrica más importante en esta problemática es la de Recall, que representa la proporción de casos positivos correctamente clasificados respecto al total de registros de la clase. Una comparación de los resultados de Recall de los tres modelos se muestra en la Figura 25.

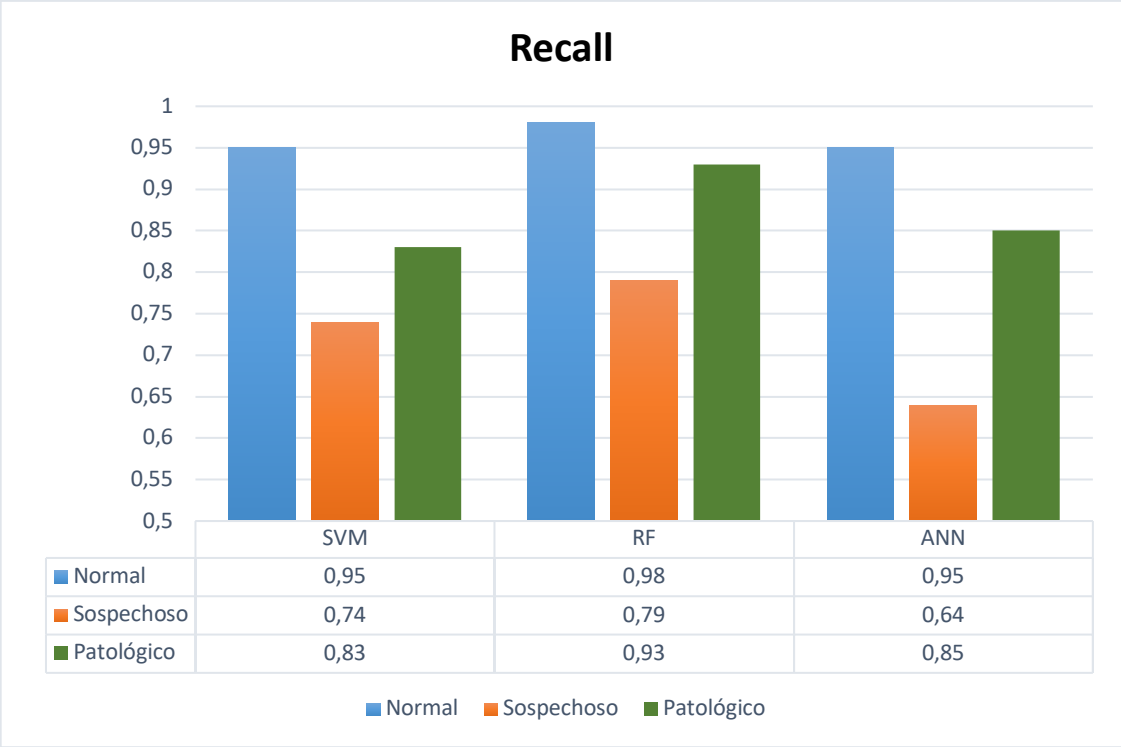


Figura 25. Comparación de los resultados de la métrica Recall de los algoritmos implementados.

El resultado de la métrica Recall para el modelo de Random Forest presenta el valor más alto en las tres clases, seguido por SVM y, por último, ANN. Se concluye que el algoritmo que presenta los mejores resultados para este problema de clasificación en concreto es Random Forest.

3.1.2 Resultados del modelo de XAI

LIME nos indica de manera descendente las características que influyen de mayor a menor medida en la decisión del modelo de aprendizaje automático. Existe un factor que indica si la variable apoya a la decisión o la contradice, en los gráficos de LIME cada característica se muestra con un color determinado. Por ejemplo, dado el caso de una instancia clasificada como “Normal”, las variables marcadas en verde son aquellas que apoyaron la clasificación de esta instancia como una perteneciente a dicha clase. Por el contrario, las variables marcadas en rojo soportan la idea de que esta instancia no debería ser clasificada como “Normal”.

El enfoque de la variable frente a la decisión del algoritmo está sujeto al rango de valores que se ubican junto al nombre de la variable. Tomando el ejemplo anterior, si la variabilidad anormal a corto plazo toma valores mayores a 61 y se muestra en color rojo, LIME indica

que este atributo soporta la idea de que esta instancia no debería ser clasificada como “Normal” en este rango, sin embargo, para valores menores a 61 la situación podría variar.

A partir de la implementación de LIME, se obtuvo un explicador que recibe una instancia y devuelve las características más influyentes en la decisión tomada por el algoritmo de aprendizaje automático. Esta sección detalla los resultados que se obtienen con distintas entradas, se ingresa una instancia por cada clase de salida que indica la salud fetal (“Normal”, “Sospechoso” y “Patológico”), específicamente.

Instancia “Normal”

Tabla 14. Valores de una instancia clasificada como "Normal".

Característica	Valor
1. baseline_value	140.000
2. accelerations	0.004
3. fetal_movement	0.001
4. uterine_contractions	0.007
5. light_decelerations	0.005
6. prolongued_decelerations	0.000
7. severe_decelerations	0.000
8. abnormal_short_ term_variability	64.000
9. mean_value_of_short_term_variability	1.100
10. percentage_of_time_with_abnormal_long_term_variability	0.000
11. mean_value_of_long_term_variability	3.800
12. histogram_mode	146.000
13. histogram_mean	140.000
14. histogram_median	144.000
15. histogram_variance	16.000

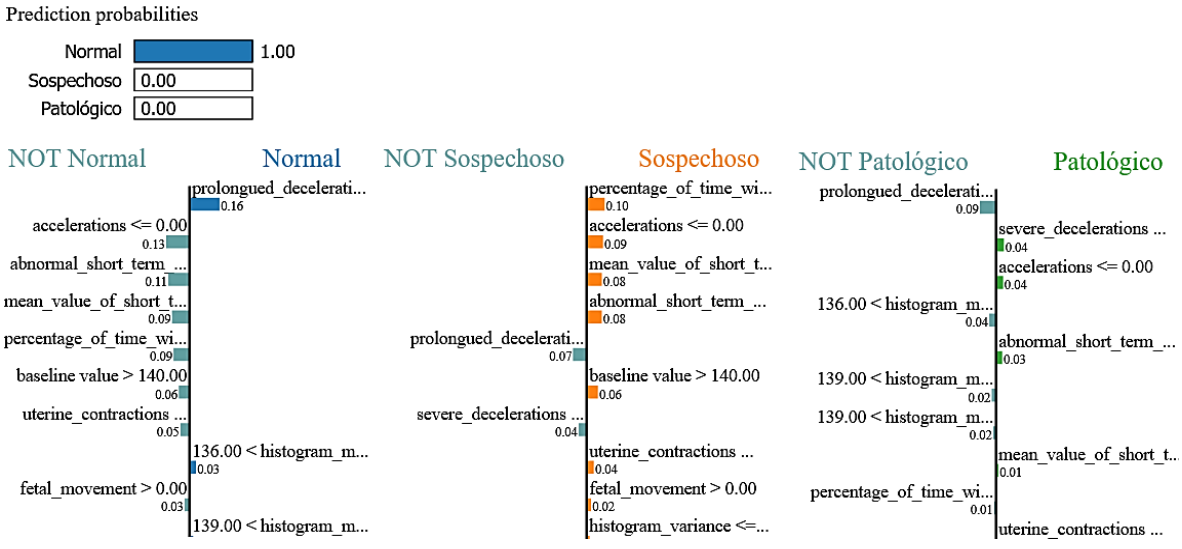


Figura 26. Explicaciones de LIME para una instancia clasificada como "Normal".

En la Tabla 14 se muestran los valores de una instancia que será clasificada como “Normal” por el algoritmo de aprendizaje automático.

Las explicaciones otorgadas por LIME para esta instancia se muestran en la Figura 26.

LIME nos da la posibilidad de visualizar con mayor detalle la clase que nos interesa, en este caso se trata de la denominada como “Normal”:

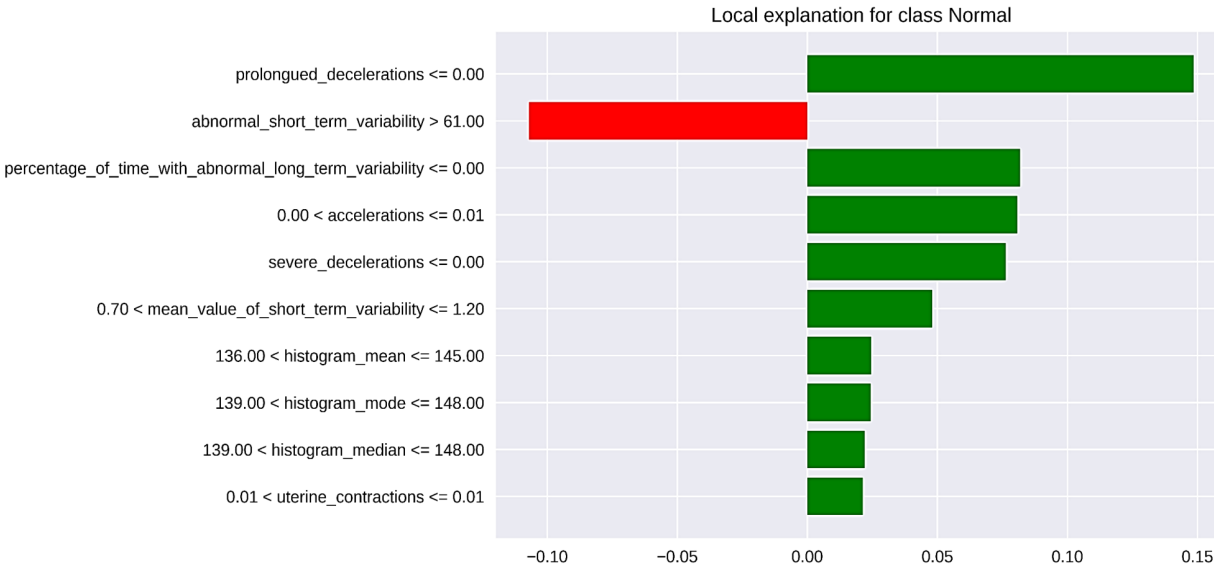


Figura 27. Explicación detallada de la clase "Normal".

Para el caso de esta instancia, LIME nos indica que las desaceleraciones prolongadas con un valor menor igual a cero aumentan la posibilidad de que la salud de un feto sea clasificada como “Normal”; por el contrario, valores superiores a 61 de la variabilidad anormal a corto plazo están en contra de esta clasificación.

Otras características que apoyan la clasificación de esta instancia como “Normal” es el porcentaje de tiempo con una variabilidad anormal a largo plazo para valores menores o iguales a cero; las aceleraciones que se encuentran en un rango entre 0 y 0.01; y las desaceleraciones severas con valores menores o iguales a 0.

Instancia “Sospechosa”

A continuación, se muestran los valores de una instancia que será clasificada como “Sospechosa” por el algoritmo de aprendizaje automático.

Tabla 15. Valores de una instancia clasificada como "Sospechosa".

Característica	Valor
1. baseline_value	145.000
2. accelerations	0.000
3. fetal_movement	0.021
4. uterine_contractions	0.000
5. light_decelerations	0.000
6. prolongued_decelerations	0.000
7. severe_decelerations	0.000
8. abnormal_short_term_variability	74.000
9. mean_value_of_short_term_variability	0.300
10. percentage_of_time_with_abnormal_long_term_variability	30.000
11. mean_value_of_long_term_variability	8.500
12. histogram_mode	145.000
13. histogram_mean	144.000
14. histogram_median	146.000
15. histogram_variance	1.000

Las explicaciones otorgadas por LIME para esta instancia se muestran a continuación:

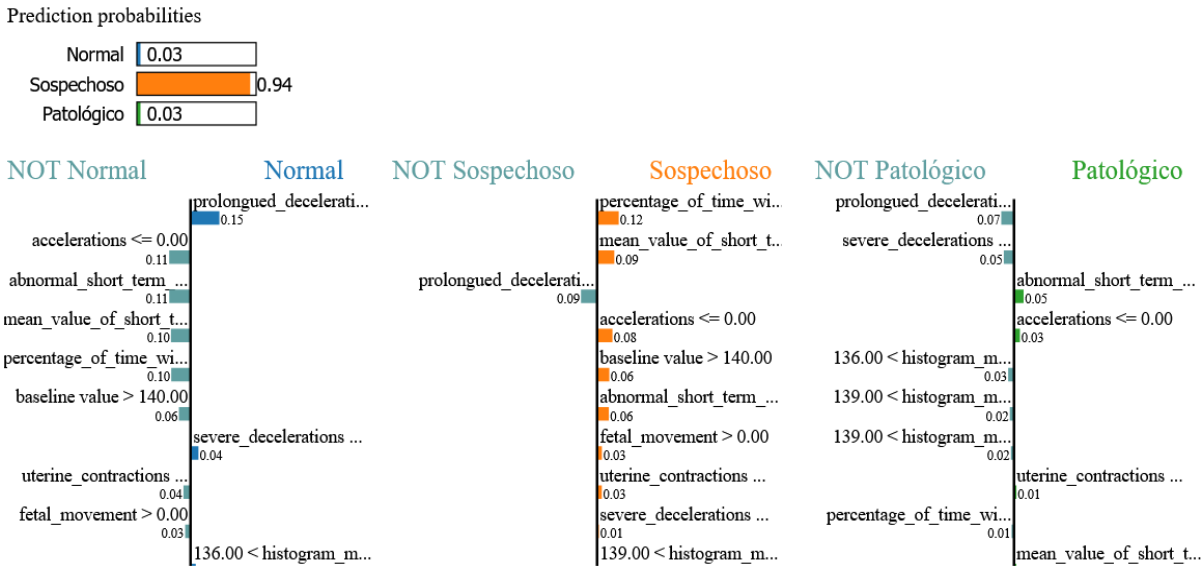


Figura 28. Explicaciones de LIME para una instancia clasificada como "Sospechosa".

La visualización en detalle de la clase “Sospechoso” se indica en la Figura 29:

Para esta instancia en específico, LIME determina que el porcentaje de tiempo con una variabilidad anormal a largo plazo para valores mayores a 10 clasifica una instancia como esta en “Sospechosa”. Además, el algoritmo indica que el valor medio de la variabilidad a corto plazo con valores menores a 0.70, las aceleraciones menores o iguales a 0, y la línea base por encima de 140, apoyan esta decisión. Por otro lado, LIME indica que una instancia

con las desaceleraciones prolongadas menores o iguales a cero tiene menos posibilidades de ser clasificada como “Sospechosa”.

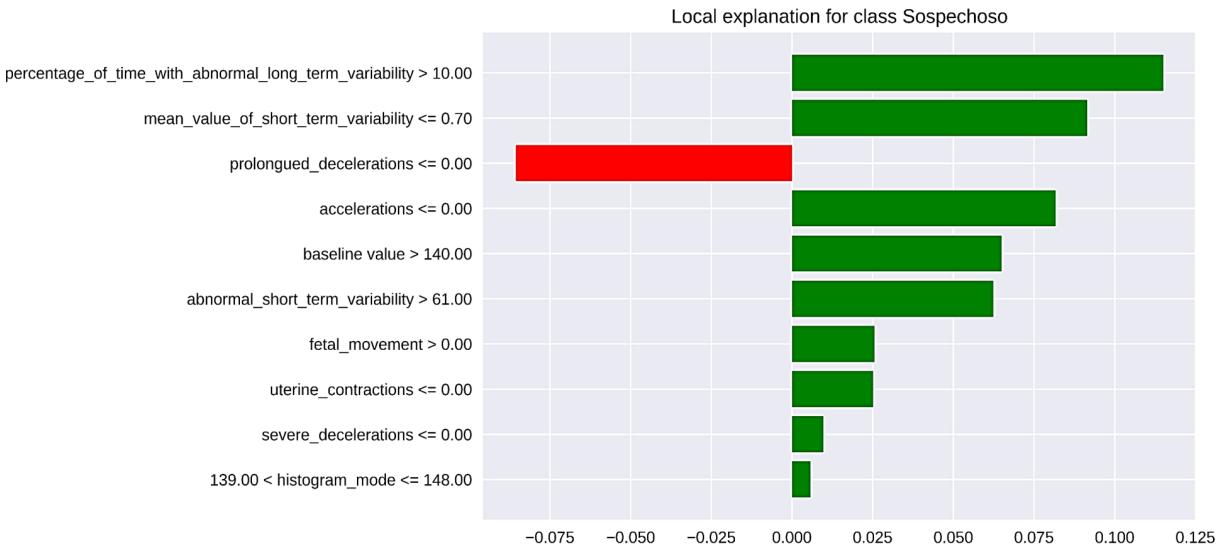


Figura 29. Explicación detallada de la clase "Sospechosa".

Instancia “Patológica”

A continuación, se muestran los valores de una instancia que será clasificada como “Patológica” por el algoritmo de aprendizaje automático.

Tabla 16. Valores de una instancia clasificada como “Patológica”.

Característica	**Valor**
1. baseline_value	133.0
2. accelerations	0.000
3. fetal_movement	0.000
4. uterine_contractions	0.000
5. light_decelerations	0.000
6. prolongued_ decelerations	0.000
7. severe_decelerations	0.000
8. abnormal_short_term_variability	73.000
9. mean_value_of_short_term_variability	0.300
10. percentage_of_time_with_abnormal_long_term_variability	84.000
11. mean_value_of_long_term_variability	3.500
12. histogram_mode	134.0
13. histogram_mean	134.0
14. histogram_median	135.0
15. histogram_variance	0.000

Las explicaciones otorgadas por LIME para esta instancia se muestran a continuación:

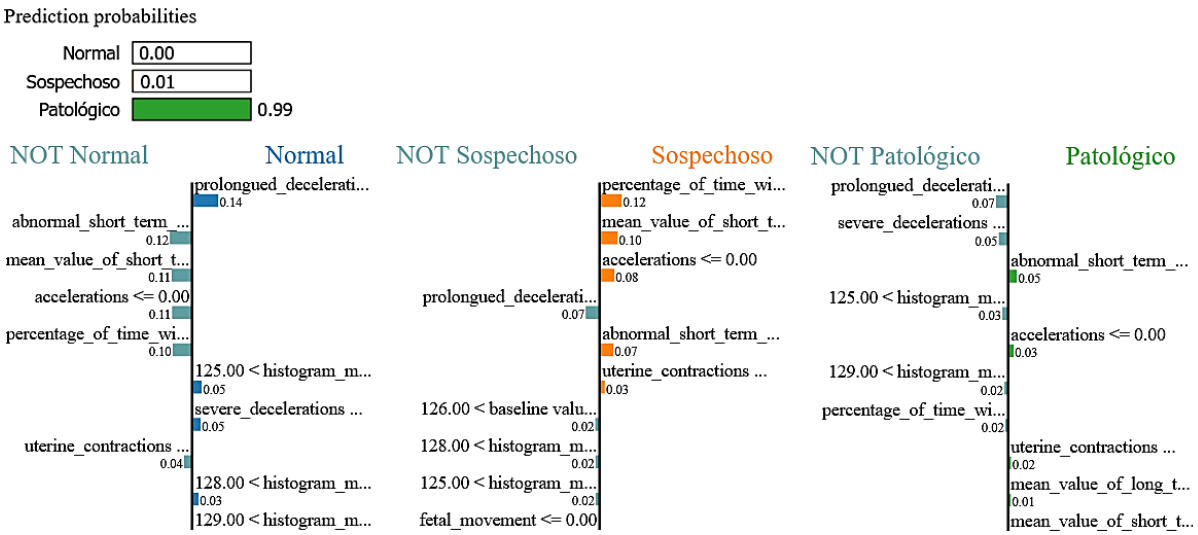


Figura 30. Explicaciones de LIME para una instancia clasificada como “Patológica”.

Visualización en detalle de la clase “Patológico”:

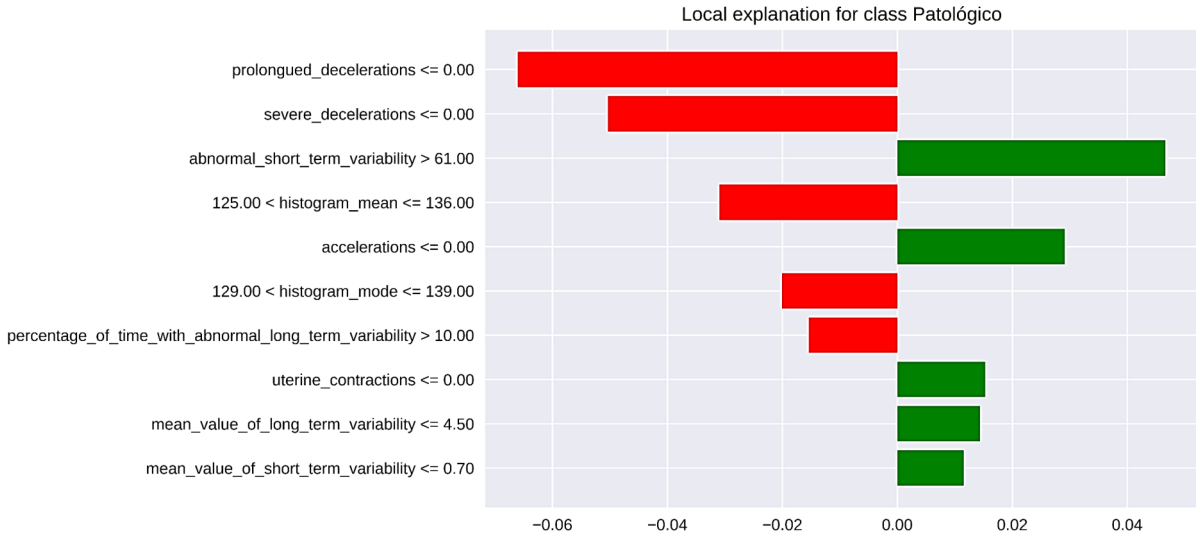


Figura 31. Explicación detallada de la clase “Patológica”.

Según LIME, las desaceleraciones prolongadas menores o iguales a 0 tienen la mayor influencia al momento de decidir si un feto se clasifica como “Patológico” o no, seguido de las desaceleraciones severas; puesto que, si una instancia posee este tipo de desaceleraciones con valores bajos, posiblemente no se trate de un feto con problemas patológicos. Además, LIME indetermina que la variabilidad anormal a corto plazo para valores mayores de 61 es un factor de peso para clasificar esta instancia como “Patológica”, similar sucede con las aceleraciones menores o iguales a 0.

3.1.3 Resultados de la evaluación del modelo de XAI

La evaluación de las explicaciones descritas anteriormente se llevó a cabo siguiendo el Modelo de Aceptación Tecnológica como se describe en la Sección 2.5. A continuación, se presentan las respuestas otorgadas por el usuario entrevistado:

Facilidad de uso percibida:

1) ¿Qué tan fácil de entender le parecieron las explicaciones obtenidas?
   En mi opinión, las explicaciones no tienen una connotación complicada, son bastante fáciles de usar y la técnica es de mi agrado. Además, los resultados del estudio se encuentran acorde con la realidad de la labor en la obstetricia. Es importante tomar en cuenta que, como profesionales en esta área, ya disponemos de una base de conocimiento que apoya la comprensión e interpretación de los resultados obtenidos; sin embargo, considero que estas explicaciones pueden llegar a ser entendidas incluso por personal no especializado en este campo en específico.
2) ¿Con qué facilidad podría implementar estas explicaciones en su trabajo?

   Las explicaciones de esta técnica se pueden transportar al momento obstétrico con sencillez; de esta forma, la toma de decisiones durante el proceso de análisis se desenvolvería de mejor manera. En mi criterio, los resultados de este estudio se pueden llevar a la realidad.

Utilidad percibida:

1) ¿Qué tan útiles le parecieron las explicaciones obtenidas?
   Las explicaciones son bastante útiles, nos permiten corroborar lo que los profesionales del área suponemos, por lo tanto, representan un apoyo en la teoría de lo que se visualiza en la realidad y sirven de evidencia para formar ciencia verdaderamente útil.
2) ¿Qué tan útil considera que sería implementar en su trabajo las explicaciones obtenidas?
   Implementar las explicaciones en mi trabajo sería totalmente útil. Disponer de herramientas como esta, que permiten tomar decisiones correctas sobre temas tan impredecibles como los que se pueden encontrar en el área de la obstetricia, podría hacer una gran diferencia. Por ejemplo, si notamos que un monitoreo durante la labor de parto no está bien, tomar la decisión correcta puede ser la diferencia entre la vida y la muerte.

Recomendación:

- ¿Qué recomendaciones daría para mejorar las explicaciones obtenidas?
  A pesar de que las explicaciones están bastante claras, en el área de la obstetricia existen elementos que podrían ser detallados con mayor profundidad, por lo que se recomendaría tratar más factores alrededor de este campo. Debido a esto, me interesaría mucho conocer un posible desarrollo futuro de esta investigación.

3.2 Conclusiones

- El presente trabajo de titulación responde a la necesidad de entendimiento en las decisiones de los algoritmos de inteligencia artificial en la clasificación de la salud fetal, para lo que se ha propuesto una comparación de algoritmos de aprendizaje automático y la implementación de un modelo demostrador de XAI a través de LIME que permita apoyar la toma de decisiones sobre diagnósticos médicos.

- El proceso de elaboración del proyecto ha sido apoyado por la metodología CRISP-DM, que provee 6 etapas claramente diferenciadas en un enfoque orientado al análisis de datos e implementación de modelos inteligentes. Los resultados de su implementación indican la falta de literatura respecto a los modelos de XAI en obstetricia; además, luego de la comparación y evaluación contextualizada de los modelos de aprendizaje automático: SVM, ANN y Random Forest, es este último algoritmo el que presenta los mejores resultados en cuanto a la clasificación de la salud fetal.

- Los resultados enfatizados por el estudio corresponden a las características influyentes en las decisiones efectuadas. En este aspecto, LIME determinó que para una instancia clasificada como “Normal”, valores bajos en aceleraciones y desaceleraciones prolongadas apoyan esta decisión. Sin embargo, altos valores de la variabilidad anormal a corto plazo pueden contradecirla. Para una instancia clasificada como “Sospechosa”, el porcentaje de tiempo prolongado con variabilidad anormal a largo plazo influye directamente en esta clasificación. Finalmente, en una instancia clasificada como “Patológica”, las desaceleraciones prolongadas y severas en valores mínimos contradicen esta decisión, por lo tanto, valores elevados podrían advertir una patología en la salud del feto. Estas explicaciones han sido validadas por un experto en el campo obstétrico a través del Modelo de Aceptación de Tecnología (TAM) en lo que respecta a la facilidad de uso y la utilidad percibida.

- El trabajo integrador ha permitido sacar a la vista algunos elementos para retener o considerar. En este caso específico, conviene no alterar la naturaleza de los datos en la fase de preparación, pues es muy común implementar técnicas de reducción de dimensionalidad o sobremuestreo de datos. Esto agregaría ruido en las explicaciones del algoritmo de XAI e impediría su correcta interpretación. Por otro lado, resulta importante considerar el contexto del problema al momento de establecer métricas de evaluación, más aún cuando se tratan temas sensibles como los hallados en la medicina.
- Si bien este trabajo representa un caso aplicable a un área específica de la salud que puede verse beneficiada por las ventajas de la AI, gran parte de su éxito depende de la calidad de los datos utilizados en el estudio. Una limitante al respecto es la naturaleza desbalanceada de los datos, no todas las características disponían de la misma cantidad de registros y se evitó aplicar un balanceo artificial con la intención de conservar la integridad de las explicaciones.
- En investigaciones futuras se plantea establecer una comparación de carácter cualitativito entre diferentes algoritmos de XAI como LIME, SHAP, XGBoost, etc., para determinar el modelo que otorga las explicaciones más útiles y fáciles de integrar para este problema de clasificación. Además, implementar las estrategias de despliegue diseñadas en este proyecto permitiría construir un sistema de información completo para el usuario final.

4 REFERENCIAS BIBLIOGRÁFICAS

[1] A. Agrawal, J. Gans, y A. Goldfarb, “What to expect from artificial intelligence”. MIT Sloan Management Review, 2017.

[2] A. F. S. Borges, F. J. B. Laurindo, M. M. Spínola, R. F. Gonçalves, y C. A. Mattos, “The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions”, *Int. J. Inf. Manage.*, vol. 57, p. 102225, 2021, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102225.

[3] H. Hagras, “Toward Human-Understandable, Explainable AI”, *Computer (Long. Beach. Calif).*, vol. 51, núm. 9, pp. 28–36, 2018, doi: 10.1109/MC.2018.3620965.

[4] C. Meske, E. Bunde, J. Schneider, y M. Gersch, “Explainable artificial intelligence: objectives, stakeholders, and future research opportunities”, *Inf. Syst. Manag.*, pp. 1–11, 2021.

[5] V. Arya *et al.*, “One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques”, *arXiv Prepr. arXiv1909.03012*, 2019.

[6] F. K. Došilović, M. Brčić, y N. Hlupić, “Explainable artificial intelligence: A survey”, en *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 210–215, doi: 10.23919/MIPRO.2018.8400040.

[7] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, y B. Yu, “Definitions, methods, and applications in interpretable machine learning”, *Proc. Natl. Acad. Sci.*, vol. 116, núm. 44, pp. 22071–22080, oct. 2019, doi: 10.1073pnas.1900654116.

[8] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, y M. A. Przybocki, “Four principles of explainable artificial intelligence”, *Gaithersburg, Maryl.*, 2020.

[9] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, y K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*, vol. 11700. Springer Nature, 2019.

[10] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, y G. Klein, “Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI”, *arXiv Prepr. arXiv1902.01876*, 2019.

[11] P. Hjortdahl, “The influence of general practitioners’ knowledge about their patients on the clinical decision-making process”, *Scand. J. Prim. Health Care*, vol. 10, núm.

4, pp. 290–294, 1992.

[12] A. Gunawardana y G. Shani, “A survey of accuracy evaluation metrics of recommendation tasks.”, *J. Mach. Learn. Res.*, vol. 10, núm. 12, 2009.

[13] A. J. London, “Artificial intelligence and black-box medical decisions: accuracy versus explainability”, *Hastings Cent. Rep.*, vol. 49, núm. 1, pp. 15–21, 2019.

[14] F. Hutter, L. Kotthoff, y J. Vanschoren, *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[15] C. Yang, A. Rangarajan, y S. Ranka, “Global model interpretation via recursive partitioning”, en *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018, pp. 1563–1570.

[16] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[17] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[18] MathWorks, “Interpretability - MATLAB & Simulink”, *What is interpretability?*, 2019. https://la.mathworks.com/discovery/interpretability.html (consultado dic. 13, 2021).

[19] M. T. Ribeiro, S. Singh, y C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier”, en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[20] B. M. Greenwell, “pdp: An R package for constructing partial dependence plots.”, *R Journal.*, vol. 9, núm. 1, p. 421, 2017.

[21] C. Molnar, G. König, B. Bischl, y G. Casalicchio, “Model-agnostic Feature Importance and Effects with Dependent Features--A Conditional Subgroup Approach”, *arXiv Prepr.*, 2020.

[22] D. F. Andrews y A. M. Herzberg, “Iris Data BT - Data: A Collection of Problems from Many Fields for the Student and Research Worker”, D. F. Andrews y A. M. Herzberg, Eds. New York, NY: Springer New York, 1985, pp. 5–8.

[23] MathWorks, “Compute Partial Dependence - MATLAB”, *Partial Dependence*, 2021.

https://la.mathworks.com/help/stats/regressiontree.partialdependence.html (consultado ene. 17, 2022).

[24] F. Träuble *et al.*, “On disentangled representations learned from correlated data”, en *International Conference on Machine Learning*, 2021, pp. 10401–10412.

[25] D. W. Apley y J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models”, *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 82, núm. 4, pp. 1059–1086, 2020.

[26] E. Algaba, V. Fragnelli, y J. Sánchez-Soriano, *Handbook of the Shapley value*. Boca Raton, FL: CRC Press, 2019.

[27] S. M. Lundberg y S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, en *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

[28] D. Showkat, “Tinkering: A Way Towards Designing Transparent Algorithmic User Interfaces”, mar. 2021.

[29] D. Garreau y U. Luxburg, “Explaining the explainer: A first theoretical analysis of LIME”, en *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1287–1296.

[30] A. Gosiewska y P. Biecek, “iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models”, mar. 2020.

[31] P. Linardatos, V. Papastefanopoulos, y S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods”, *Entropy* , vol. 23, núm. 1. 2021, doi: 10.3390/e23010018.

[32] R. L. Marchese Robinson, A. Palczewska, J. Palczewski, y N. Kidley, “Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets”, *J. Chem. Inf. Model.*, vol. 57, núm. 8, pp. 1773–1792, ago. 2017, doi: 10.1021/acs.jcim.6b00753.

[33] H. T. Shen, Y. Zhu, W. Zheng, y X. Zhu, “Half-Quadratic Minimization for Unsupervised Feature Selection on Incomplete Data”, *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, núm. 7, pp. 3122–3135, 2021, doi: 10.1109/TNNLS.2020.3009632.

[34] S. Albawi, T. A. Mohammed, y S. Al-Zawi, “Understanding of a convolutional neural

network", en *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[35] S. Li, N. Marsaglia, C. Garth, J. Woodring, J. Clyne, y H. Childs, "Data reduction techniques for simulation, visualization and data analysis", en *Computer Graphics Forum*, 2018, vol. 37, núm. 6, pp. 422–447.

[36] J. Dieber y S. Kirrane, "Why model why? Assessing the strengths and limitations of LIME", *CoRR*, vol. abs/2012.0, 2020, [En línea]. Disponible en: https://arxiv.org/abs/2012.00093.

[37] S. Kaur *et al.*, "Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives", *IEEE Access*, vol. 8, pp. 228049–228069, 2020, doi: 10.1109/ACCESS.2020.3042273.

[38] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, y V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease", *Futur. Gener. Comput. Syst.*, vol. 83, pp. 366–373, 2018, doi: https://doi.org/10.1016/j.future.2018.02.009.

[39] W. Yue, Z. Wang, H. Chen, A. Payne, y X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", *Designs* , vol. 2, núm. 2. 2018, doi: 10.3390/designs2020013.

[40] Z. Liao, D. Li, X. Wang, L. Li, y Q. Zou, "Cancer Diagnosis Through IsomiR Expression with Machine Learning Method", *Current Bioinformatics*, vol. 13, núm. 1. pp. 57–63, 2018, doi: http://dx.doi.org/10.2174/1574893611666160609081155.

[41] T. Richter, B. Fishbain, E. Fruchter, G. Richter-Levin, y H. Okon-Singer, "Machine learning-based diagnosis support system for differentiating between clinical anxiety and depression disorders", *J. Psychiatr. Res.*, vol. 141, pp. 199–205, 2021, doi: https://doi.org/10.1016/j.jpsychires.2021.06.044.

[42] K. S. Betts, S. Kisely, y R. Alati, "Predicting common maternal postpartum complications: leveraging health administrative data and machine learning", *BJOG An Int. J. Obstet. Gynaecol.*, vol. 126, núm. 6, pp. 702–709, may 2019, doi: https://doi.org/10.1111/1471-0528.15607.

[43] K. K. Venkatesh *et al.*, "Machine Learning and Statistical Models to Predict Postpartum Hemorrhage.", *Obstet. Gynecol.*, vol. 135, núm. 4, pp. 935–944, abr. 2020, doi: 10.1097/AOG.0000000000003759.

[44] A. I. Naimi, R. W. Platt, y J. C. Larkin, “Machine Learning for Fetal Growth Prediction”, *Epidemiology*, vol. 29, núm. 2, 2018, [En línea]. Disponible en: https://journals.lww.com/epidem/Fulltext/2018/03000/Machine_Learning_for_Fetal_Growth_Prediction.17.aspx.

[45] J. Spilka *et al.*, “Using nonlinear features for fetal heart rate classification”, *Biomed. Signal Process. Control*, vol. 7, núm. 4, pp. 350–357, 2012, doi: https://doi.org/10.1016/j.bspc.2011.06.008.

[46] J. H. Miao y K. H. Miao, “Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classification”, *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, núm. 5, 2018, doi: 10.14569/IJACSA.2018.090501.

[47] A. Akbulut, E. Ertugrul, y V. Topcu, “Fetal health status prediction based on maternal clinical history using machine learning techniques”, *Comput. Methods Programs Biomed.*, vol. 163, pp. 87–100, 2018, doi: https://doi.org/10.1016/j.cmpb.2018.06.010.

[48] Y. Lu, X. Zhang, X. Fu, F. Chen, y K. K. L. Wong, “Ensemble Machine Learning for Estimating Fetal Weight at Varying Gestational Age”, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, núm. 01 SE-IAAI Technical Track: Emerging Papers, pp. 9522–9527, jul. 2019, doi: 10.1609/aaai.v33i01.33019522.

[49] L. J. G. D. S. M. W. C. A. Devane D y V. Smith, “Cardiotocography versus intermittent auscultation of fetal heart on admission to labour ward for assessment of fetal wellbeing”, *Cochrane Database Syst. Rev.*, núm. 1, 2017, doi: 10.1002/14651858.CD005122.pub5.

[50] D. D. Alfirevic Z y G. M. L. Gyte, “Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour”, *Cochrane Database Syst. Rev.*, núm. 5, 2013, doi: 10.1002/14651858.CD006066.pub2.

[51] A. Z. G. G. M. L. Grivell RM y D. Devane, “Antenatal cardiotocography for fetal assessment”, *Cochrane Database Syst. Rev.*, núm. 9, 2015, doi: 10.1002/14651858.CD007863.pub4.

[52] G. Improta, M. Romano, A. M. Ponsiglione, P. Bifulco, G. Faiella, y M. Cesarelli, “Computerized Cardiotocography: A Software to Generate Synthetic Signals”, *J. Heal. Med. Informatics*, vol. 05, 2014.

[53] R. Wirth y J. Hipp, “CRISP-DM: Towards a standard process model for data mining”,

Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. Data Min., ene. 2000.

[54] C. Schröer, F. Kruse, y J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model", *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: https://doi.org/10.1016/j.procs.2021.01.199.

[55] A. Nadali, E. N. Kakhky, y H. E. Nosratabadi, "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system", en *2011 3rd International Conference on Electronics Computer Technology*, 2011, vol. 6, pp. 161–165, doi: 10.1109/ICECTECH.2011.5942073.

[56] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, y S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review", *Inf. Softw. Technol.*, vol. 51, núm. 1, pp. 7–15, 2009.

[57] S. Banerjee *et al.*, "Deep Relational Reasoning for the Prediction of Language Impairment and Postoperative Seizure Outcome Using Preoperative DWI Connectome Data of Children With Focal Epilepsy", *IEEE Trans. Med. Imaging*, vol. 40, núm. 3, pp. 793–804, 2021, doi: 10.1109/TMI.2020.3036933.

[58] M. Kiani, J. Andreu-Perez, H. Hagras, M. L. Filippetti, y S. Rigato, "A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Developmental Neuroscience", en *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020, pp. 1–8, doi: 10.1109/FUZZ48607.2020.9177711.

[59] W. S. Liew, C. K. Loo, y S. Wermter, "Emotion Recognition Using Explainable Genetically Optimized Fuzzy ART Ensembles", *IEEE Access*, vol. 9, pp. 61513–61531, 2021, doi: 10.1109/ACCESS.2021.3072120.

[60] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, y T. Wilaiprasitporn, "Parkinson's Disease Recognition Using SPECT Image and Interpretable AI: A Tutorial", *IEEE Sens. J.*, vol. 21, núm. 20, pp. 22304–22316, oct. 2021, doi: 10.1109/JSEN.2021.3077949.

[61] A. Raison, P. Bourdon, C. Habas, y D. Helbert, "Explicability in resting-state fMRI for gender classification", en *2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME)*, oct. 2021, pp. 5–8, doi: 10.1109/CVPRW53098.2021.00199.

[62] D. O. Nahmias y K. L. Kontson, "Easy Perturbation EEG Algorithm for Spectral Importance (EasyPEASI): A Simple Method to Identify Important Spectral Features

of EEG in Deep Learning Models", en *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2398–2406, doi: 10.1145/3394486.3403289.

[63] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, y C. Fookes, "A Robust Interpretable Deep Learning Classifier for Heart Anomaly Detection Without Segmentation", *IEEE J. Biomed. Heal. Informatics*, vol. 25, núm. 6, pp. 2162–2171, 2021, doi: 10.1109/JBHI.2020.3027910.

[64] S. Ghosh, P. Tino, y K. Bunte, "Visualisation and knowledge discovery from interpretable models", en *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206702.

[65] P. A. Moreno-Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees", en *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 4902–4910, doi: 10.1109/BigData50022.2020.9378460.

[66] J. Duell, X. Fan, B. Burnett, G. Aarts, y S.-M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records", en *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4, doi: 10.1109/BHI50953.2021.9508618.

[67] N. Potie, S. Giannoukakos, M. Hackenberg, y A. Fernandez, "On the Need of Interpretability for Biomedical Applications: Using Fuzzy Models for Lung Cancer Prediction with Liquid Biopsy", en *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6, doi: 10.1109/FUZZ-IEEE.2019.8858976.

[68] A. Kumar, R. Manikandan, U. Kose, D. Gupta, y S. C. Satapathy, "Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis", *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, núm. 3s, oct. 2021, doi: 10.1145/3457187.

[69] C. Panigutti, A. Perotti, y D. Pedreschi, "Doctor XAI: An Ontology-Based Approach to Black-Box Sequential Data Classification Explanations", en *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 629–639, doi: 10.1145/3351095.3372855.

[70] N. Keller, M. A. Jenny, C. A. Spies, y S. M. Herzog, "Augmenting Decision Competence in Healthcare Using AI-based Cognitive Models", en *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, nov. 2020, pp. 1–4, doi:

10.1109/ICHI48887.2020.9374376.

[71] A. Tahmassebi, J. Martin, A. Meyer-Baese, y A. H. Gandomi, “An Interpretable Deep Learning Framework for Health Monitoring Systems: A Case Study of Eye State Detection using EEG Signals”, en *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 211–218, doi: 10.1109/SSCI47803.2020.9308230.

[72] G. Dong, Y. Ma, y A. Basu, “Feature-Guided CNN for Denoising Images From Portable Ultrasound Devices”, *IEEE Access*, vol. 9, pp. 28272–28281, 2021, doi: 10.1109/ACCESS.2021.3059003.

[73] C. K. Leung, D. L. x. Fung, D. Mai, Q. Wen, J. Tran, y J. Souza, “Explainable Data Analytics for Disease and Healthcare Informatics”, en *25th International Database Engineering & Applications Symposium*, 2021, pp. 65–74, doi: 10.1145/3472163.3472175.

[74] Q. Ye, J. Xia, y G. Yang, “Explainable AI for COVID-19 CT Classifiers: An Initial Comparison Study”, en *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 521–526, doi: 10.1109/CBMS52027.2021.00103.

[75] D. R. Chittajallu *et al.*, “XAI-CBIR: Explainable AI System for Content based Retrieval of Video Frames from Minimally Invasive Surgery Videos”, en *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 66–69, doi: 10.1109/ISBI.2019.8759428.

[76] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, y K. H. Ryu, “Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction”, *IEEE Access*, vol. 9, pp. 123672–123688, 2021, doi: 10.1109/ACCESS.2021.3110336.

[77] M. R. Karim *et al.*, “DeepKneeExplainer: Explainable Knee Osteoarthritis Diagnosis From Radiographs and Magnetic Resonance Imaging”, *IEEE Access*, vol. 9, pp. 39757–39780, 2021, doi: 10.1109/ACCESS.2021.3062493.

[78] P. F. Khan y K. Meehan, “Diabetes prognosis using white-box machine learning framework for interpretability of results”, en *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 2021, pp. 1501–1506, doi: 10.1109/CCWC51732.2021.9375927.

[79] J. Kim, M. Kim, y Y. M. Ro, “Interpretation of Lesional Detection via Counterfactual

Generation", en *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 96–100, doi: 10.1109/ICIP42928.2021.9506282.

[80] N. Seedat, V. Aharonson, y Y. Hamzany, "Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning", en *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6, doi: 10.1109/CSDE50874.2020.9411529.

[81] K. A. Thakoor, S. C. Koorathota, D. C. Hood, y P. Sajda, "Robust and Interpretable Convolutional Neural Networks to Detect Glaucoma in Optical Coherence Tomography Images", *IEEE Trans. Biomed. Eng.*, vol. 68, núm. 8, pp. 2456–2466, 2021, doi: 10.1109/TBME.2020.3043215.

[82] F. Stieler, F. Rabe, y B. Bauer, "Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach", en *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1802–1809, doi: 10.1109/CVPRW53098.2021.00199.

[83] U. Pawar, C. T. Culbert, y R. O'Reilly, "Evaluating Hierarchical Medical Workflows using Feature Importance", en *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 265–270, doi: 10.1109/CBMS52027.2021.00075.

[84] D. Dua y C. Graff, "UCI Machine Learning Repository". 2017, [En línea]. Disponible en: http://archive.ics.uci.edu/ml.

[85] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, y L. Pereira-Leite, "Sisporto 2.0: A program for automated analysis of cardiotocograms", *J. Matern. Fetal. Med.*, vol. 9, núm. 5, pp. 311–318, sep. 2000, doi: https://doi.org/10.1002/1520-6661(200009/10)9:5<311::AID-MFM12>3.0.CO;2-9.

[86] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python", *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[87] N. Cristianini y E. Ricci, "Support Vector Machines BT - Encyclopedia of Algorithms", M.-Y. Kao, Ed. Boston, MA: Springer US, 2008, pp. 928–932.

[88] A. Liaw y M. Wiener, "Classification and Regression by randomForest", *R News*, vol. 2, núm. 3, pp. 18–22, 2002, [En línea]. Disponible en: https://cran.r-project.org/doc/Rnews/.

[89] D. Buschmann, C. Enslin, H. Elser, D. Lütticke, y R. H. Schmitt, "Data-driven decision

support for process quality improvements", *Procedia CIRP*, vol. 99, pp. 313–318, 2021, doi: https://doi.org/10.1016/j.procir.2021.03.047.

[90] W. S. McCulloch y W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *Bull. Math. Biophys.*, vol. 5, núm. 4, pp. 115–133, 1943.

[91] Martín~Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems". 2015, doi: 10.5281/zenodo.4724125.

[92] M. Roodschild, J. Gotay Sardiñas, y A. Will, "A new approach for the vanishing gradient problem on sigmoid activation", *Prog. Artif. Intell.*, vol. 9, núm. 4, pp. 351–360, 2020, doi: 10.1007/s13748-020-00218-y.

[93] F. Davis y F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", *MIS Q.*, vol. 13, p. 319, sep. 1989, doi: 10.2307/249008.

5 ANEXOS

ANEXO I

Enlace dirigido al código fuente desarrollado en este trabajo de titulación:

https://colab.research.google.com/drive/1GP-E2sHWeJ7rHIWQmUdl-lOjrVvGv5lT?usp=sharing

ANEXO II

Enlace dirigido a la matriz resultante de la extracción de información de la revisión sistemática de literatura:

https://epnecuador-my.sharepoint.com/:x:/g/personal/ismael_rivas_epn_edu_ec/EUEk8XM7C1lFk4CmdZwV4JMBmyP311Y3MBq7pBjV649mlg?rtime=qHd1fPT72Ug

ANEXO III

Enlace dirigido a la matriz resultante de los algoritmos de AI:

https://epnecuador-my.sharepoint.com/:x:/g/personal/ismael_rivas_epn_edu_ec/EeJpj766LyBJp2smcm1ro-kBBVk1yXoWNih0ERToqSGzOw?e=SgWMyv

ANEXO IV

Enlace dirigido a la transcripción de la entrevista efectuada al profesional de la salud que evaluó el algoritmo de XAI:

https://epnecuador-my.sharepoint.com/:b:/g/personal/ismael_rivas_epn_edu_ec/EST9noJ-SwREkDNTFNOYPJQBUUS5JR_RDojh68kB4hYqOw?e=CR8P2P