

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 SHAP 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**


BRYAN DAVID ORTUÑO BARRERA

DIRECTOR: EDISON FERNANDO LOZA AGUIRRE

DMQ, febrero 2022

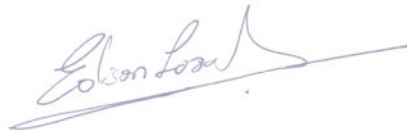
CERTIFICACIONES

Yo, BRYAN DAVID ORTUÑO BARRERA 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.



BRYAN DAVID ORTUÑO BARRERA

Certifico que el presente trabajo de integración curricular fue desarrollado por BRYAN DAVID ORTUÑO BARRERA, bajo mi supervisión.



EDISON FERNANDO LOZA AGUIRRE

Certificamos que revisamos el presente trabajo de integración curricular.

NOMBRE_REVISOR1
REVISOR1 DEL TRABAJO DE
INTEGRACIÓN CURRICULAR

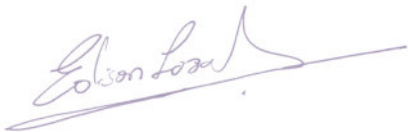
NOMBRE_REVISOR2
REVISOR2 DEL TRABAJO DE
INTEGRACIÓN CURRICULAR

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.



BRYAN DAVID ORTUÑO BARRERA



EDISON FERNANDO LOZA AGUIRRE

DEDICATORIA

Realizo una especial dedicatoria de todo este esfuerzo realizado a mi madre Marieta Barrera Jaramillo y mi padre Wilson Bolívar Ortuño Chauca, quienes me han apoyado en todo momento, han sabido guiarme y hacer de mi una persona mejor. A mis hermanos, sobrinos y cuñada que con su ejemplo y ayuda incondicional me han motivado a seguir adelante y no rendirme. A mis compañeros de la Escuela Politécnica Nacional con quienes he seguido este largo camino de aprendizaje, apoyándonos por superar cada obstáculo y con quienes he vivido anécdotas extraordinarias.

También me gustaría realizar una dedicatoria personal que demuestra lo lejos que puedo llegar y las cosas tan grandes que puedo lograr.

AGRADECIMIENTO

Agradezco de todo corazón a mi madre Marieta Barrera Jaramillo y a mi padre Wilson Bolívar Ortuño Chauca quienes me han dado la oportunidad de estudiar y poder llegar a este punto de mi vida académica. Gran parte de este esfuerzo es en un su honor ya que sin ellos esta investigación presentada como trabajo de integración curricular no hubiera sido posible realizarla.

A mi tutor de tesis el PhD. Edison Fernando Loza Aguirre quien me ha sabido demostrar que ser una excelente persona es primordial para ser un buen profesional, quien con su conocimiento me ha guiado a realizar esta investigación con éxito, agradecerle por todo su apoyo y confianza tanto en lo académico como en lo profesional.

Agradezco a la Escuela Politécnica Nacional donde me he forjado como un estudiante y profesional de excelencia, como solo en esta prestigiosa universidad lo hubiera podido imaginar y por la cual me he esforzado por seguir este camino de retos y mucho aprendizaje.

Finalmente me gustaría agradecer a la Dra. Tatiana Tapia especialista en el área de obstetricia, la cual me ha sabido apoyar con su tiempo y conocimientos para realizar la validación de esta investigación desde una perspectiva profesional de la medicina.

ÍNDICE DE CONTENIDO

CERTIFICACIONES	I
DECLARACIÓN DE AUTORÍA	III
DEDICATORIA	IV
AGRADECIMIENTO	V
ÍNDICE DE CONTENIDO	VI
ÍNDICE DE TABLAS	VII
ÍNDICE DE FIGURAS	VIII
RESUMEN	X
ABSTRACT	XI
1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO	1
1.1 Objetivo general	2
1.2 Objetivos específicos	2
1.3 Alcance	2
1.4 Marco teórico	3
1.4.1 Inteligencia Artificial Explicable.....	3
1.4.2 Shapley Additive Explanations (SHAP)	7
1.4.3 AI en el área de Obstetricia	8
2 METODOLOGÍA	13
2.1 Descripción General de la metodología CRISP-DM	13
2.2 Aplicación de la metodología CRISP-DM	15
2.2.1 Comprensión del negocio	15
2.2.2 Comprensión de datos.....	22
2.2.3 Preparación de datos.....	25
2.2.4 Modelado	34
2.2.5 Despliegue	38
3 PRUEBAS, RESULTADOS, CONCLUSIONES Y RECOMENDACIONES ...	39
3.1 Pruebas	39
3.2 Resultados	40
3.2.1 Modelos de AI	40
3.2.2 Modelo de XAI	43
3.2.3 Evaluación médica.....	51
3.3 Conclusiones	53
3.4 Recomendaciones	54
4 REFERENCIAS BIBLIOGRÁFICAS	55
5 ANEXOS	63

ANEXO I	64
ANEXO II	65
ANEXO III	66
ANEXO IV	67

Índice de Tablas

<i>Tabla 1: Artículos totales encontrados para realizar la SLR</i>	17
<i>Tabla 2: Número de artículos totales luego del primer filtro basado en el abstract</i>	18
<i>Tabla 3: Número de artículos restantes luego de la selección basada en el contenido de cada artículo</i>	18
<i>Tabla 4: Clasificación de artículos acorde a cada área de medicina</i>	21
<i>Tabla 5: Algoritmos de AI más usados en el conjunto de artículos</i>	21
<i>Tabla 6: Algoritmos de XAI más usados en el conjunto de artículos</i>	21
<i>Tabla 7: Conjunto de datos clasificados por su naturaleza</i>	22
<i>Tabla 8: Descripción del conjunto de datos [44]</i>	25
<i>Tabla 9: Resultados de la ejecución de los diferentes algoritmos de AI con las clases Normal (N), Sospechoso (S) y Patológico (P)</i>	42
<i>Tabla 10: Comparación Recall de los modelos de AI</i>	42
<i>Tabla 11: Instancia de la clase Normal</i>	44
<i>Tabla 12: Instancia de la clase Sospechosa</i>	45
<i>Tabla 13: Instancia de la clase Patológica</i>	45
<i>Tabla 14: Valores medios de cada atributo en el conjunto de datos</i>	46

Índice de Figuras

<i>Figura 1 Modelo de AI que aprendido a usar la nieve en el fondo como un indicador de "lobo". [5]</i>	<i>4</i>
<i>Figura 2: Salida explicativa de LIME en un modelo de Regresión Logística [17].....</i>	<i>6</i>
<i>Figura 3 Procedimiento LRP [19]</i>	<i>6</i>
<i>Figura 4 Valores de Shap [23]. Predicción $f(x)$ explicada por la suma ϕ_i de los efectos de cada variable</i>	<i>8</i>
<i>Figura 5 Cardiotocografía y análisis de la forma de onda ST de un feto con acidosis metabólica al nacer</i>	<i>9</i>
<i>Figura 6: Aceleraciones en una cardiotocografía, imagen obtenida de geekymedics.com</i> <i>11</i>	<i>11</i>
<i>Figura 7: Metodología CRISP-DM, Figura adaptada de [28].....</i>	<i>13</i>
<i>Figura 8: Diseño SLR</i>	<i>15</i>
<i>Figura 9: Diagrama de flujo SLR, imagen modificada de [30].....</i>	<i>19</i>
<i>Figura 10: Detalle del conjunto de datos SisPorto cardiotocograms dataset.....</i>	<i>25</i>
<i>Figura 11: Muestra categórica de la clasificación de la salud fetal en el conjunto de datos</i> <i>26</i>	<i>26</i>
<i>Figura 12: Distribución de datos parte 1.....</i>	<i>27</i>
<i>Figura 13: Distribución de datos parte 2.....</i>	<i>27</i>
<i>Figura 14: Matriz de correlación del conjunto de datos.....</i>	<i>28</i>
<i>Figura 15: Ejemplo gráfico de alta varianza y alto sesgo. Imagen obtenida de: towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.....</i>	<i>30</i>
<i>Figura 16 Ejemplo gráfico de características relevantes, redundantes e irrelevantes [46]30</i>	<i>30</i>
<i>Figura 17: Árbol de decisión para escoger el método de selección de categoría, Imagen obtenida de MachineLearningMastery.com.....</i>	<i>31</i>
<i>Figura 18: Selección de características del histograma usando el algoritmo SelectKBest32</i>	<i>32</i>
<i>Figura 19: Bloxpot que indica la varianza en la distribución de datos de cada atributo....</i>	<i>32</i>
<i>Figura 20: Boxplot que indica la varianza en la distribución de los datos de cada atributo luego de realizar una estandarización mediante el algoritmo StandarScaler de la librería sklearn</i>	<i>33</i>
<i>Figura 21: Red Neuronal Artificial para evaluar la salud fetal</i>	<i>35</i>
<i>Figura 22: Matriz de confusión</i>	<i>36</i>
<i>Figura 23: Matriz de confusión del modelo SVM.....</i>	<i>40</i>
<i>Figura 24: Matriz de confusión del modelo RF.....</i>	<i>41</i>
<i>Figura 25: Matriz de confusión del modelo ANN.....</i>	<i>41</i>
<i>Figura 26: Gráfico de decisión para realizar explicaciones de la instancia Normal</i>	<i>47</i>

<i>Figura 27: Gráfico de cascada para realizar explicaciones de la instancia Normal</i>	<i>47</i>
<i>Figura 28: Gráfico de decisión para realizar explicaciones de la instancia Sospechosa ..</i>	<i>48</i>
<i>Figura 29: Gráfico de cascada para realizar explicaciones de la instancia Sospechosa..</i>	<i>49</i>
<i>Figura 30: Gráfico de decisión para realizar explicaciones de la instancia Patológica</i>	<i>49</i>
<i>Figura 31: Gráfico de cascada para realizar explicaciones de la instancia Patológica</i>	<i>50</i>

RESUMEN

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. Sin embargo La AI se presenta como una “caja negra” que esconde el proceso de aprendizaje para la generación de resultados.

En esta investigación presentada como trabajo de titulación curricular se realiza la aplicación del algoritmo de Inteligencia Artificial Explicada (XAI, por sus siglas en inglés) SHAP como método explicativo para la evaluación de la salud fetal. Para lo cual se utiliza los datos de cardiocografías obtenidos del repositorio UCI Machine Learning para la aplicación de AI en el área de obstetricia.

Para aplicar el método de XAI primero se realizó una selección de un algoritmo de AI para la cual se evaluaron 3 métodos de aprendizaje automático muy populares como son Suport Vector Machine, Random Forest y una Red Neuronal Artificial. Siendo Random Forest el algoritmo que obtuvo los mejores resultados en la clasificación de la salud fetal.

Los resultados han sido evaluados con un especialista del área de obstetricia quién ha validado la capacidad explicativa del algoritmo SHAP y ha encontrado en este método una ayuda para interpretar las salidas de un algoritmo de AI lo cual se considera favorable para su implementación en la medicina, específicamente en el área de obstetricia.

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

ABSTRACT

In recent decades, medicine has found in artificial intelligence (AI) a driver in the search for solutions. Thus, through machine learning algorithms, tasks as specific as patient diagnosis are favored by their ability to predict based on data. However, AI is presented as a "black box" that hides the learning process for the generation of results.

In this research presented as a curricular degree work, the application of the Explained Artificial Intelligence (XAI) SHAP algorithm is carried out as an explanatory method for the evaluation of fetal health. For which the cardiocography data obtained from the UCI Machine Learning repository is used for the application of AI in the area of obstetrics.

To apply the automatic XAI method, a selection of an AI algorithm was first made, for which 3 very popular learning methods were evaluated, such as Support Vector Machine, Random Forest and an Artificial Neural Network. Being Random Forest the algorithm that obtained the best results in the classification of fetal health.

The results have been evaluated with a specialist in the area of obstetrics who has validated the explanatory capacity of the SHAP algorithm and has found in this method a help to interpret the outputs of an AI algorithm, which is considered favorable for its implementation in medicine. specifically in the area of obstetrics.

KEYWORDS: XAI, SHAP, decision making, 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 comprometedor, 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 en este sentido 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 la implementación de 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 SHAP para el apoyo en la toma de decisiones médicas para la salud fetal.

1.2 Objetivos específicos

- Realizar una Revisión Sistemática de Literatura referente a la XAI y sus implicaciones en la medicina.
- Efectuar una evaluación comparativa entre algoritmos de aprendizaje automático para la predicción de la salud fetal.
- Desarrollar un modelo de XAI de tipo SHAP para soportar la toma de decisiones médicas en la salud fetal.
- Validar la capacidad explicativa de los algoritmos de XAI 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 de la salud fetal. Para ello, 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 de aprendizaje de máquina, se construirá el modelo de XAI con SHAP. 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 e 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

El término de XAI fue acuñado por primera vez en 2004 por Van Lent et al. [1] con el fin de describir la capacidad de un sistema para explicar el comportamiento de entidades controladas por algoritmos de AI en aplicaciones de juegos simulados. XAI tiene como objetivo describir la lógica detrás de los algoritmos tradicionales de IA, haciendo que la toma de decisiones sea un proceso que incluso las personas con pocas habilidades informáticas pueden comprender [2]. Además, comprender su funcionamiento puede ser uno de los indicadores clave para evaluar el modelo y mantener un equilibrio entre complejidad e interpretabilidad [3].

A menudo no solo es relevante la precisión de un modelo, es necesario que sea preciso y explicable [4], esto también aumentaría su aceptación social y de AI en general. El campo de la XAI está destinado a dar explicaciones de los modelos de caja negra de la AI [5] en el cual el algoritmo toma millones de puntos de datos como entradas y correlaciona ciertas características de los datos para producir una salida. Este proceso es en gran medida autodirigido y difícil de interpretar para los científicos de datos, programadores y usuarios en general.

El sesgo de AI puede introducirse en los algoritmos, por ejemplo, por sesgos conscientes o inconscientes de los desarrolladores, o puede infiltrarse a través de errores no detectados [6]. En cualquier caso, los resultados de un algoritmo sesgado están distorsionados, posiblemente de una manera que sea ofensiva para las personas que utilizan dicho sistema.

Un ejemplo de esto se puede evidenciar en el artículo de Ribeiro et al. 2016 [5], en el cual plantea el problema de clasificación "¿Husky o Wolf?", en el que algunos perros Husky fueron clasificados como lobos en imágenes, es decir, la clasificación era errónea. Si se observan los píxeles que más influyeron en la decisión en la Figura 1, se puede notar que la decisión del algoritmo solo depende del fondo, por lo que el modelo había aprendido a usar la nieve en el fondo como un indicador de "lobo":

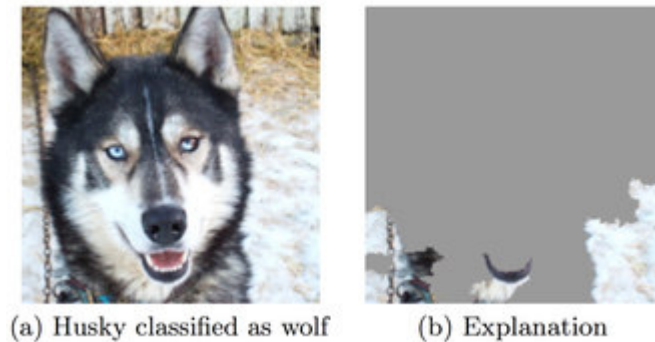


Figura 1 Modelo de AI que aprendido a usar la nieve en el fondo como un indicador de "lobo". [5]

Como tal, XAI es un conjunto de procesos y métodos que permiten a las personas comprender y confiar en los resultados generados por los algoritmos de aprendizaje automático [5]. XAI a su vez, se utiliza para describir un modelo de AI, su efecto esperado y el posible sesgo de los errores sistemáticos; ayudando a caracterizar las precisiones del modelo, la transparencia y los resultados en la toma de decisiones respaldada por la AI [7].

Existen principalmente dos enfoques para la aplicación de XAI [8]: El primero consiste en la realización de explicaciones con modelos de AI simples que tienden a dar resultados fáciles de entender y por otro lado tenemos los métodos post hoc, los cuales se realizan posterior a la ejecución de un algoritmo de AI y sirven para realizar una interpretación de sus resultados los cuales no son comprensibles incluso para los usuarios que se encuentran en el campo de la informática [9].

Modelos interpretables

Generalmente los modelos simples de AI son modelos interpretables [10]. La idea en estos modelos es cuantificar directamente el cálculo y los parámetros; y con esto mantenerlos a un nivel interpretable. Estos modelos son muy conocidos en el campo de la ciencia de datos, por ejemplo, las regresiones lineales, los árboles de decisión y los bosques

aleatorios interpretables [5]. Aquí, por ejemplo, la varianza explicable de una regresión lineal se utiliza para comprender los factores que influyen en los resultados.

Sin embargo, estos modelos simples tienen desventajas [10]. Se puede considerar a la más clara como el hecho de sufrir mala calidad predictiva debido a su simpleza. Los modelos lineales se limitan a las relaciones lineales, por lo que generalmente fallan cuando las predicciones tienden a efectos no lineales [11].

Post-Hoc XAI: explicación de los modelos de caja negra

Una opción mejorada para XAI es utilizar un modelo de AI de caja negra complejo, pero más preciso. A continuación, se puede estudiar la comprensión del resultado y realizar una explicación posterior a su entrenamiento. Dichos métodos se denominan métodos post-hoc porque se aplican después de que el modelo ha sido entrenado[12].

El desafío de Post-Hoc XAI es hacer que un modelo de caja negra sea cuantificable retrospectivamente [13]. Se utilizan varios métodos, que se registran durante el entrenamiento o, por ejemplo, revisan todo el modelo nuevamente para cuantificarlo. Algunos de los métodos más comunes para realizar explicaciones post-hoc son los siguientes [11], [12], [14]–[16]:

LIME: "Local Interpretable Model-Agnostic Explanations", tienen la autoafirmación de hacer que todos los modelos sean explicables. La idea principal de LIME es hacer que un modelo existente de AI sea comprensible para una persona con pocos conocimientos informáticos. LIME tiene el propósito de actuar sin conocimiento de un modelo específico (agnóstico del modelo), se puede utilizar actualmente con datos tabulares, imágenes y texto, siempre y cuando LIME pueda crear variaciones a partir de la entrada.

La salida del modelo debe ser posible en forma de una clase o valor especificado para que LIME pueda evaluar los resultados de la entrada modificada. En la Figura 2 podemos observar un ejemplo de la salida de LIME en un estudio realizado por Dieber J. 2020 [17] en el cual evalúa la probabilidad de que llueva, siendo 0.83 la probabilidad de lluvia y marcado de color azul los factores que más influyen en la decisión del modelo. Junto a cada factor determinante de la decisión se puede observar un valor de influencia. Por ejemplo, en los datos de Dieber J. se puede observar que la velocidad de la ráfaga de viento (WIndGustSpeed) tiene la mayor puntuación, por lo cual se puede determinar que es el valor más determinante para predecir la probabilidad de lluvia [17].

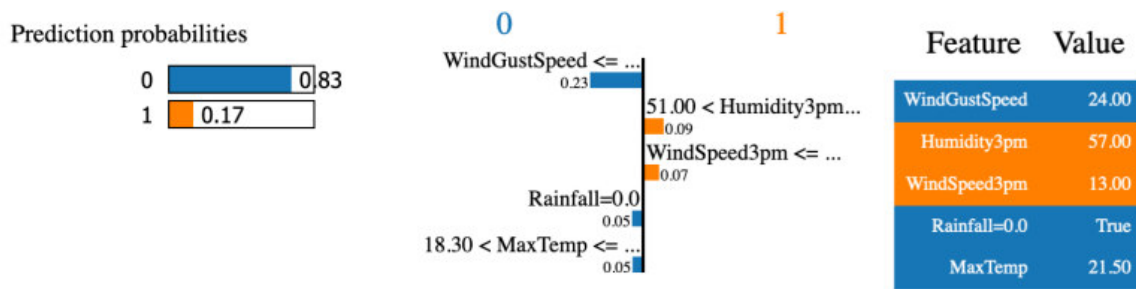


Figura 2: Salida explicativa de LIME en un modelo de Regresión Logística [17]

Método contrafactual: El Método contrafactual utiliza el hecho de que la salida de un modelo es el resultado directo de la entrada para hacer que la AI sea explicable [18]. En términos concretos, esto significa que los elementos de entrada, por ejemplo, un atributo o una imagen se manipulan hasta que se puede observar un cambio en la salida. Si este método se repite sistemáticamente, es posible determinar qué sutilezas en la entrada explican la salida.

Propagación de relevancia por capas (LRP): Mientras que el método contrafactual manipula la entrada, LRP intenta garantizar la explicabilidad a través de la retro propagación, es decir, la distribución hacia atrás [18], como se puede observar en la Figura 3. Para este propósito, la salida se rastrea hasta los nodos ponderados de la capa en una red neuronal. Esto permite identificar las combinaciones de nodo-borde más importantes y así marcar la mayor influencia de ciertas partes de la entrada.

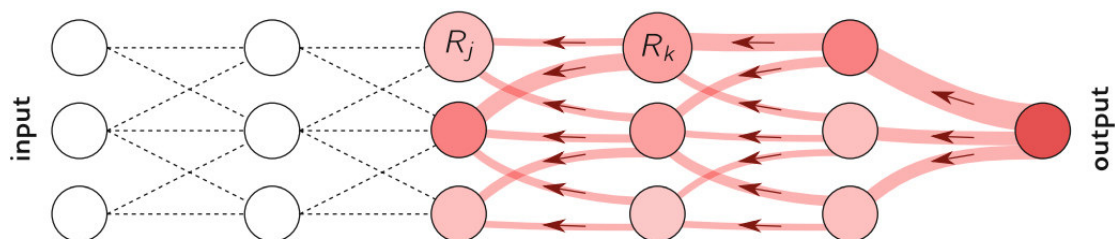


Figura 3 Procedimiento LRP [19]

Racionalización: Particularmente son de gran interés también los enfoques en los que las máquinas de caja negra (como los robots) pueden explicar su comportamiento por sí mismas [20]. Para realizar esta acción, se necesitan cálculos más profundos para registrar

las razones por las que se desencadenan acciones y hacer que esta información sea interpretable para las personas.

Otros métodos: Además de estos métodos bien conocidos, existen otros métodos para hacer la AI Explicable, por ejemplo, Expectativa condicional individual (ICE), Efectos locales acumulados (ALE), Interacción de características, Importancia de la característica de permutación, Sustitutos globales, Reglas de alcance, Explicaciones aditivas de Shapley (SHAP) y algunos más [11], [16], [21].

1.4.2 Shapley Additive Explanations (SHAP)

Valores Shapley

El Valor Shapley es un concepto de solución de la teoría de juegos cooperativos [21]. El método que sirve para encontrar los valores Shapley fue propuesto por Lloyd Shapley en el año 1953 [22] con el objetivo de medir la contribución de cada jugador en un juego cooperativo [16]. Con este fin, se plantea la idea que “ n ” jugadores que participan de manera colectiva con el propósito de obtener una recompensa “ p ” la cual se debe repartir entre los “ n ” jugadores acorde a la contribución individual que realizó cada uno; La cual se considera como el valor de Shapley [21].

Para calcular el valor de Shapley de un jugador, se deben considerar todos los subconjuntos del conjunto de todos los jugadores que contienen ese jugador. Para cada uno de estos subconjuntos, el beneficio con él (generalmente: el valor de la alianza) se resta del beneficio sin él. Esta diferencia se calcula con el factor $\frac{(s-1)!(n-s)!}{n!}$ donde n es el número de todos los jugadores y s representa el número de jugadores del subconjunto en cuestión. Shapley propuso el valor de un juego representado por (N, v) que se da para cada jugador $i \in N$ mediante la siguiente formula [22]:

$$\varphi_i(v) = \sum_{s \subseteq N: i \in s} \frac{(s-1)!(n-s)!}{n!} [v(s) - v(s-i)] \text{ donde } n = |N| \text{ y } s = |S|$$

El valor de Shapley se puede interpretar como la contribución marginal que se espera del jugador i , es decir, el valor promedio de las contribuciones marginales $[v(s) - v(s-i)]$ del jugador a todas las alianzas no vacías $S \in 2^N$, considerando que el tamaño de la alianza del jugador ($1 \leq s \leq n$) es equiprobables y todas las coaliciones de tamaño S tienen la misma probabilidad [22].

Metodología SHAP

La metodología SHAP fue acuñada por primera vez en 2017 por Lundberg S y Lee S [4] es un enfoque de teoría de juegos que combina valores de Shapley con modelos de regresión lineal local. Los valores de Shapley se basan en la idea de que el resultado de la predicción se divide equitativamente entre todas las características y que se deben considerar todas las combinaciones de características posibles para determinar la importancia de una sola característica [21].

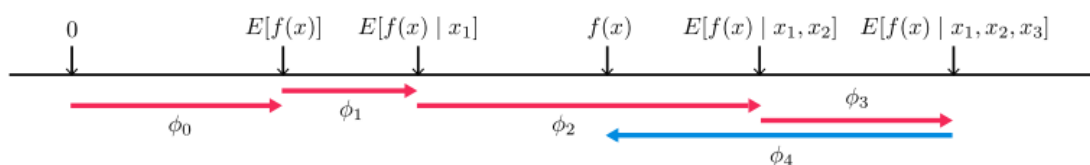


Figura 4 Valores de Shap [23]. Predicción $f(x)$ explicada por la suma ϕ_i de los efectos de cada variable

El valor de Shap computacionalmente es muy caro de calcular, lo cual se considera una desventaja de esta metodología. Debido a esto, se han desarrollado diferentes métodos de estimación como KernelSHAP (método inspirado en LIME) o TreeSHAP (método basado en árboles de decisión). KernelSHAP es un enfoque de shapley lento que teóricamente es agnóstico al modelo, pero rara vez es utilizado por su complejidad computacional [24]. En cambio, el enfoque de TreeSHAP utiliza la construcción de árboles para reducir los costos computacionales. El algoritmo recursivo guardará en memoria todos los conjuntos posibles de variables que descienden en cada hoja del árbol, y para cada ejemplo se utilizan los valores de las hojas [23].

TreeSHAP se calcula en tiempo polinomial, no en tiempo exponencial. La idea básica es mover todos los subconjuntos posibles simultáneamente. Para calcular predicciones para un solo árbol, debe realizar un seguimiento de la cantidad de subconjuntos para cada nodo de decisión. Depende del subconjunto del nodo principal y de las capacidades de división.

1.4.3 AI en el área de Obstetricia

La AI es un campo de investigación en rápido desarrollo que ofrece amplias perspectivas. Sus aplicaciones médicas, particularmente en ginecología y obstetricia, podrían mejorar la atención al paciente y la calidad de la atención.

Obstetricia

La monitorización de la Frecuencia Cardíaca Fetal (FHR, por sus siglas en inglés) se ha convertido en una parte indispensable del embarazo y, lo que es más importante, la evaluación fetal durante el parto [25]. La FHR se refiere entonces al control de la frecuencia cardíaca fetal y las contracciones uterinas (CU). Estas dos señales forman lo que también se conoce como la cardiotocografía fetal (CTG). Como principal fuente de información para los fetos que obviamente no son aptos para la observación directa, la monitorización CTG proporciona a los obstetras información sobre la salud del feto [26].

En obstetricia, la AI ha sido aplicada en la evaluación de la frecuencia cardíaca fetal (FHR) durante el parto: a pesar de décadas de progreso, este monitoreo conserva cierta subjetividad. En la actualidad, existen pocos estudios, pero algunos resultados indican que el uso de algoritmos automatizados podría clasificar correctamente el FHR con una sensibilidad del 72% al 94%, y una especificidad del 78% al 91% [25].

Cardiotocografía

CTG se define como el registro gráfico de la frecuencia cardíaca fetal y las contracciones uterinas mediante el uso de dispositivos electrónicos indicados para la evaluación de la condición fetal [26].

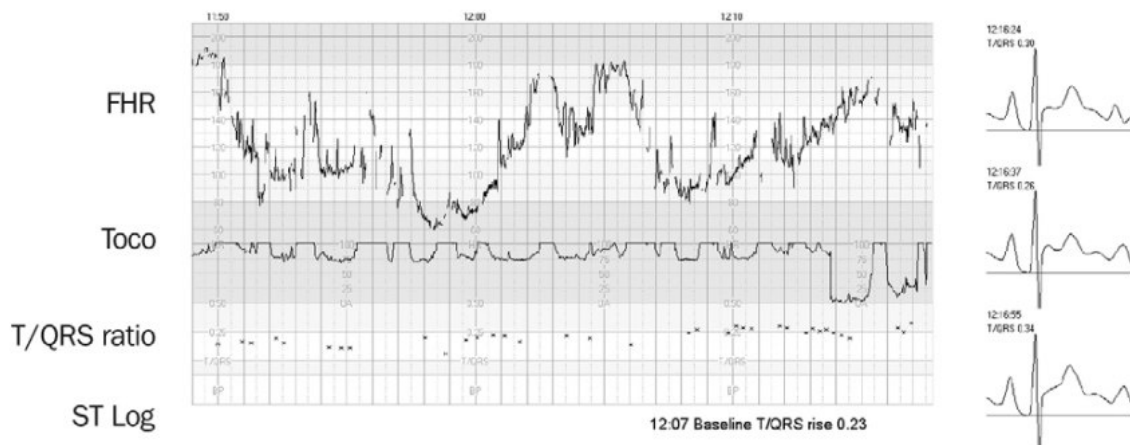


Figura 5 Cardiotocografía y análisis de la forma de onda ST de un feto con acidosis metabólica al nacer

En la Figura 5 [26] se muestra la cardiotocografía de un feto que tuvo acidosis metabólica al nacer. El panel izquierdo muestra la frecuencia cardíaca fetal (FHR), la actividad uterina (Toco) y la relación T / QRS con declaraciones de registro ST. El CTG muestra valores iniciales fluctuantes, variabilidad mantenida de la frecuencia cardíaca fetal y desaceleraciones variables. La relación T / QRS aumenta constantemente a $> 0,25$. El panel derecho muestra el segmento ST elevado y las relaciones T / QRS entre $0 \cdot 26$ y $0 \cdot 34$. El bebé nació con fórceps por sufrimiento fetal con un pH de la arteria del cordón umbilical de 6,88, déficit de bases de $14 \cdot 9$ mmol / L, y una puntuación de Apgar de 4 a 1 min y 8 a 5 min [26].

Variables de una cardiotocografía

La evaluación de los resultados de CTG se logra mediante la evaluación de sus tres componentes principales: frecuencia basal (predominante), variabilidad y cambios a corto plazo (aceleraciones y desaceleraciones).

La frecuencia basal (BF) es la frecuencia instantánea (FC fetal en un momento dado) predominante más común en CTG. Se refiere al nivel medio de FC fetal alrededor del cual tienen lugar las alteraciones de frecuencia instantánea.

La frecuencia basal se evalúa entre "las brechas" de los cambios a corto plazo. El BF normal de un feto oscila entre 110 y 150 latidos/min (lpm). La taquicardia fetal se diagnostica una vez que $BF \geq 150$ b / min; mientras que la bradicardia fetal una vez que $BF \leq 110$ b/min.

Variabilidad

La duración del ciclo cardíaco es diferente para cada feto y cambia constantemente. Estos cambios se denominan variabilidad de frecuencia instantánea o simplemente "variabilidad". La variabilidad se compone de dos componentes variabilidad a corto y largo plazo [27]:

La variabilidad a corto plazo, que también se conoce como variabilidad latido a latido, se compone de cambios en la FC fetal entre cada ciclo cardíaco adyacente. Estos cambios son menores, los cuales van de 1 a 5 lpm. Debido a estos cambios, la curva CTG tiene una apariencia más irregular que recta.

La variabilidad a largo plazo se compone de ondas de aceleración o desaceleración (oscilaciones), que varían de varias a varias decenas de ciclos cardíacos.

Una oscilación constituye toda una onda de aceleración y desaceleración con respecto al BF. El recuento de oscilación normal oscila entre 3-6 lpm con una amplitud de 6-15 lpm y rara vez lo supera.

Ambos componentes de variabilidad están inextricablemente vinculados entre sí. Forman un circuito de ritmo cardíaco coherente: Una expresión numerosa de la amplitud de oscilación de variabilidad a término tardío es la siguiente: **variabilidad normal:** 6-15 lpm; **disminuida:** 3-5 lpm; **ausente:** la amplitud se vuelve indetectable; **aumentada:** 25 lpm.

A diferencia de la frecuencia basal y la variabilidad que permanecen constantes durante algún tiempo, la duración máxima de los cambios a corto plazo es de aproximadamente 10 min. Tanto el tiempo de aceleración como el de desaceleración son muy diferentes, desde varios segundos hasta varios minutos con una amplitud diferente (10 a 30-50 t./min.)

Aceleraciones y desaceleraciones

Las aceleraciones se definen como un aumento corto y rápido de la FC fetal (Figura 6). Es la respuesta cardíaca a los movimientos fetales, también conocida como el "Reflejo Miocárdico", caracterizado por un aumento de la FC en respuesta a la actividad física [26].

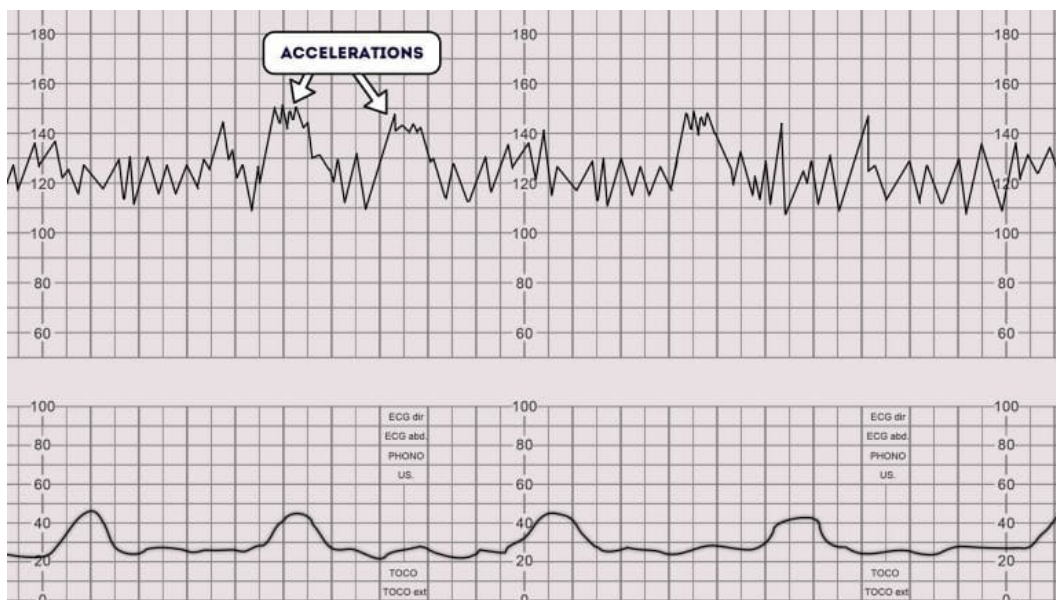


Figura 6: Aceleraciones en una cardiotocografía, imagen obtenida de geekymedics.com

Las desaceleraciones se definen como una disminución a corto plazo en la FC fetal. Las desaceleraciones, que ocurren después de cada contracción uterina, se consideran desaceleraciones periódicas. Existen diferentes tipos de desaceleraciones [27]:

Las desaceleraciones tempranas son uniformes, en el cual su inicio y cese es equivalente al comienzo y al final de la contracción uterina. Este tipo de desaceleraciones pueden deberse a un aumento de la presión intracraneal fetal (PIC) durante la contracción uterina.

Las desaceleraciones tardías representan una situación clínica grave, a pesar de su profundidad. Estos son particularmente peligrosos si se registran después de cada contracción uterina con una variabilidad reducida o ausente.

Las desaceleraciones variables son polimórficas y su inicio no es igual con respecto al inicio de la contracción uterina. Esas son las desaceleraciones registradas más comunes (> 80%) durante el trabajo de parto caracterizadas por una rápida disminución de la FC fetal en varias decenas de lpm seguida de un retorno más rápido o lento a la línea de base.

Las desaceleraciones prolongadas son aquellas que tienden a durar más de 2 min e indican una anomalía en la FC fetal.

2 METODOLOGÍA

2.1 Descripción General de la metodología CRISP-DM

Esta investigación será guiada por la metodología CRISP-DM [28] para el cumplimiento de los objetivos específicos propuestos. Las fases que integran esta metodología son las siguientes (Figura 7):

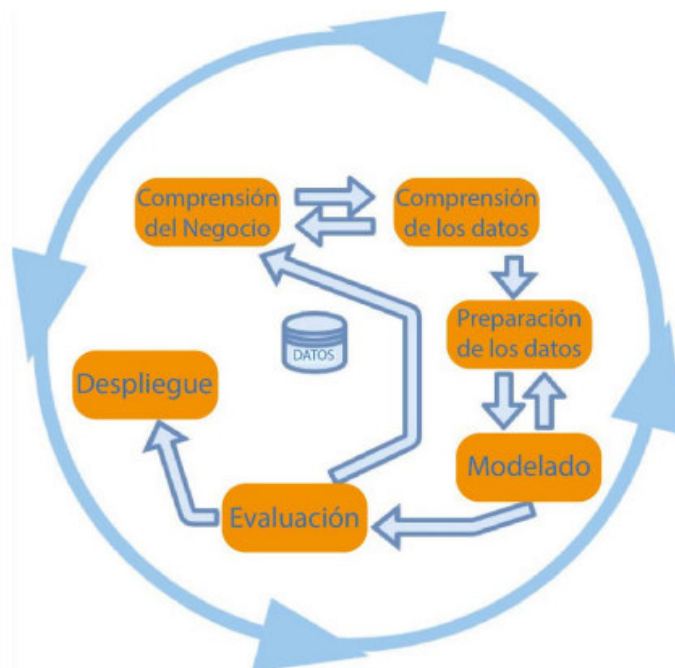


Figura 7: Metodología CRISP-DM, Figura adaptada de [28]

1. Comprensión del negocio:

La fase de comprensión del negocio consiste en establecer objetivos y requisitos específicos para la minería de datos que guiarán el proyecto. Es esencial evaluar la situación: determinar la disponibilidad de recursos, los requisitos, los riesgos y las contingencias; mientras se tienen presentes las necesidades del cliente o el público beneficiado con el proyecto.

2. Comprensión de datos:

Como parte de la comprensión de datos, se intenta obtener una descripción general de los datos disponibles. 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 siguiente fase de modelado. Esta puede ser una tarea larga en algunos casos, pues consiste en corregir, reemplazar o eliminar valores erróneos o faltantes como parte de una limpieza minuciosa de los datos.

4. Modelado:

Como regla general, se pueden utilizar varias técnicas de modelado de minería de datos para un problema específico. Algunas técnicas imponen exigencias especiales a la estructura de datos, esto puede significar que es necesario dar un paso atrás a 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 su implementación.

5. Evaluación:

En esta fase debe evaluarse si 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. El proyecto de minería de datos se evalúa retrospectivamente, se determina si se han considerado todos los factores importantes y si sus atributos 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 cliente 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.

2.2 Aplicación de la metodología CRISP-DM

2.2.1 Comprensión del negocio

En un primer paso se determinarán los objetivos de negocio identificando las necesidades de los profesionales de la salud para mejorar el diagnóstico de la salud fetal. Para realizar la comprensión del negocio en el presente trabajo de titulación se realiza una revisión sistemática de literatura (SLR) para entender los estudios que se realizan sobre las aplicaciones de XAI en el campo de la medicina.

Revisión Sistemática de Literatura SLR

Para esta sección se utilizó la metodología de Kitchenham [29] para la elaboración de Revisiones sistemáticas de Literatura que se orientan en la ingeniería de software; La cual nos permite identificar los objetivos y evaluar la literatura relevante sobre el tema de XAI en la medicina, sus implicaciones y estudios relevantes que aporten información importante en la investigación de XAI en la salud fetal. En la Figura 8 se puede observar el diseño planteado la elaboración de la SLR en esta sección.

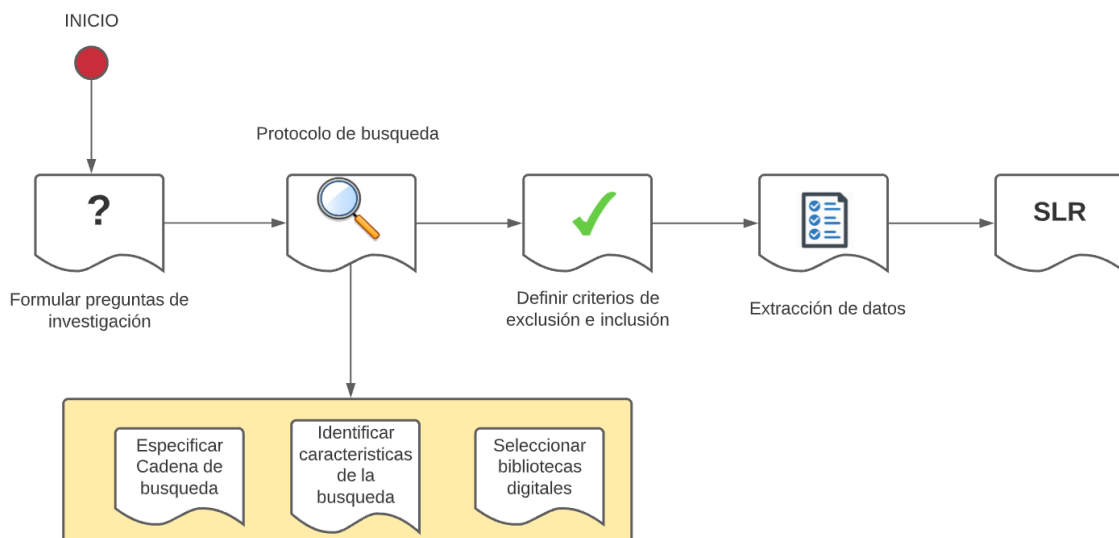


Figura 8: Diseño SLR

Preguntas de investigación

Se han identificado las siguientes preguntas para la SLR que busca respaldar esta investigación:

1. ¿En qué áreas de medicina se han realizado investigaciones de XAI?
2. ¿Qué problemas de la medicina se resuelven actualmente con XAI?
3. ¿Qué algoritmos de AI y XAI se implementan en los estudios encontrados?
4. ¿Cuál es la naturaleza de los datos que se utilizan para los estudios encontrados?
5. ¿Qué técnicas de validación se utilizan para evaluar los modelos de AI y XAI?

Search String (SS)

Se realizó una búsqueda en dos bases de datos muy conocidas como son "ACM library" e "IEEE Xplore".

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

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)

La cadena de búsqueda tiene características que nos ayudan a obtener una cantidad considerable de artículos de gran interés y que tienen gran relación con el tema de investigación. En primera instancia se realiza una búsqueda de artículos científicos que hayan realizado una experimentación de XAI, la siguiente característica importante en nuestra investigación es aquellos artículos que hablen de XAI también traten temas médicos por lo que se especifica medicina en todas las secciones del artículo para encontrar la mayor cantidad de artículos relacionados y específicos de esta investigación.

Criterios de inclusión y exclusión

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.

Artículos encontrados mediante la SS

El número de artículos encontrados en la búsqueda está expresado en la Tabla 1.

	Número de artículos
IEEE EXPLORE	92
ACM	32

Tabla 1: Artículos totales encontrados para realizar la SLR

Con el fin de obtener los artículos que aporten más información a la investigación se han planteado filtros de inclusión, lo cual nos ayuda a reducir aquellos que sean irrelevantes a la investigación. Con este objetivo se plantea eliminar aquellos artículos que realicen explicaciones sin utilizar metodologías de XAI, a manera de una explicación de datos, también se incluye la eliminación de aquellos artículos que no realicen experimentación de XAI, es decir, artículos teóricos con fundamentos de XAI.

Selección de artículos basado en su Abstract

No todos los artículos que se obtienen con la cadena de búsqueda cuentan con información relevante para la investigación, por esta razón, surge la necesidad realizar una serie de filtros que nos ayuden a seleccionar los artículos más importantes. En primera instancia se realizó una evaluación del resumen de cada artículo para analizar si el artículo cumple con los filtros de inclusión y aporta información relevante a la investigación de XAI en medicina.

En este primer filtro basado en el resumen de cada artículo pudimos obtener una reducción considerable del total. En el caso de aquellos obtenidos de la biblioteca IEEE Explore se redujeron 54 artículos los cuales representan un 58.69% del total, mientras que de los obtenidos de la biblioteca de ACM se obtuvieron 24 artículos menos, los cuales representan el 75% del total de artículos.

	Número de artículos
IEEE EXPLORE	39
ACM	8

Tabla 2: Número de artículos totales luego del primer filtro basado en el abstract

Selección de artículos basado en su contenido

No es posible asegurar que los artículos restantes, luego de realizarse el primer filtro, cuenten con un estudio relevante a la investigación de XAI en el área de medicina, esto debido a que solo se ha realizado una evaluación de su resumen; por lo que consecuente al primer filtro se realiza una revisión de todo el contenido del artículo y con esto se determinar si realmente es un artículo importante para el campo de XAI en medicina.

En el segundo filtro basado en el contenido de cada artículo se obtuvo una segunda reducción, en este caso menos significativa a la anterior. Para los artículos obtenidos de la biblioteca IEEE Explore se eliminaron 16 artículos, los cuales representan el 42.10% del total luego del primer filtro de selección y representa una reducción del 76.08% del total de artículos obtenidos por la cadena de búsqueda.

En el caso de los artículos obtenidos de la biblioteca ACM Explore se obtuvo una reducción de 4 artículos, los cuales representan el 50% del total obtenido luego del primero filtro de selección y con lo cual representa una reducción del 87.5% del total de artículos obtenido por la cadena de búsqueda.

	Número de artículos
IEEE EXPLORE	23
ACM	4

Tabla 3: Número de artículos restantes luego de la selección basada en el contenido de cada artículo

El flujo de selección de artículos destacados para la revisión sistemática de literatura se puede visualizar en la Figura 9 la cual realiza una representación de la obtención de artículos iniciales a los cuales se les realizó un proceso de filtración para finalmente obtener los artículos que han cumplido con los criterios de inclusión aceptados para la presente investigación.

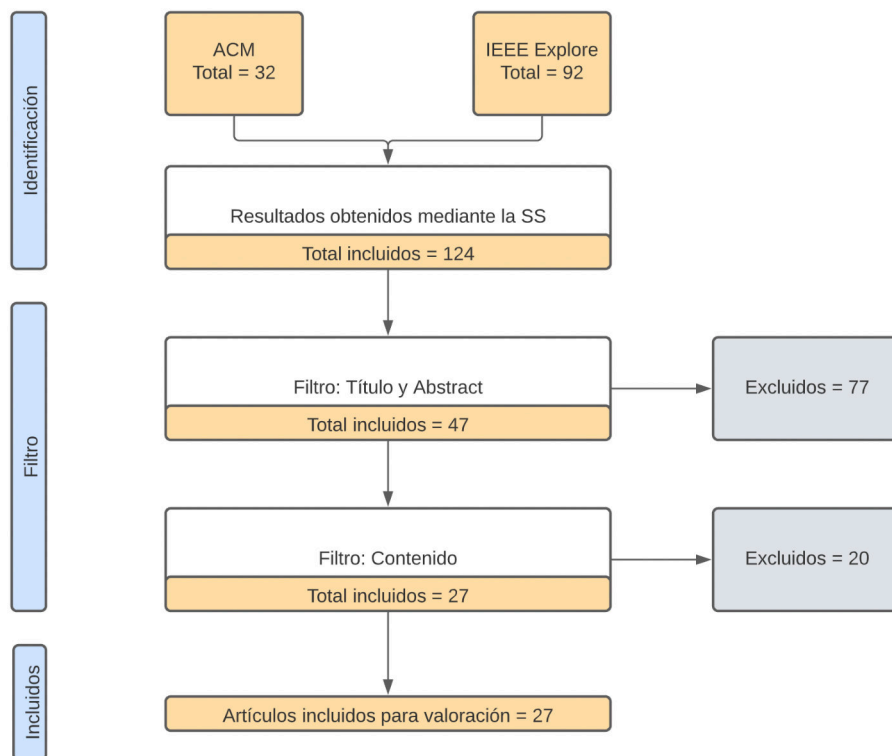


Figura 9: Diagrama de flujo SLR, imagen modificada de [30]

Extracción de información

La Tabla 3 contiene los resultados numéricos de artículos una vez realizados los filtros, por lo tanto, estos son utilizados para realizar una extracción general de información. Acorde a las preguntas de investigación previamente planteadas se ha realizado la extracción de los siguientes atributos de cada artículo:

- Área de medicina en la que se aplica el estudio
- Problema que busca resolver
- Solución o propuesta planteada
- Algoritmos de AI y de XAI utilizados
- Naturaleza del conjunto de datos usados
- Métodos de validación aplicados

Los resultados totales que se han extraído de los 27 artículos se encuentran en el Anexo 2. Con el fin de obtener una clasificación de los artículos obtenidos se ha realizado una clasificación cuantitativa acorde a los siguientes aspectos:

- Áreas de medicina en los que se aplican estos estudios
- Algoritmos de AI y de XAI utilizados
- Naturaleza del conjunto de datos utilizados

Áreas de medicina

El primer enfoque de extracción de datos cuantitativo se basa en conocer las áreas de medicina en las que se han realizado aplicaciones de XAI. Para esto, se obtiene una clasificación del número de artículo obtenidos en las diferentes áreas médicas encontradas.

Área medicina	Referencia	Número de artículos
Neurología	[31], [32], [33], [34], [35], [36]	6
Educación quirúrgica	[37]	1
Diferentes áreas de la medicina	[38]	1
Cardiología	[39], [40], [41]	3
Procesamiento de imágenes biomédicas	[42], [43]	2
Oncología	[42], [45], [44]	3
Reumatología	[46]	1
Gestión hospitalaria	[47], [48], [49]	3
Endocrinología	[50]	1
Neumología	[51]	1

Odontología	[52]	1
Oftalmología	[53]	1
Dermatología	[54]	1
Epidemiología	[55], [56]	2

Tabla 4: Clasificación de artículos acorde a cada área de medicina

Algoritmos de AI y de XAI utilizados

El segundo enfoque de extracción de información cuantitativa nos ayuda a conocer los algoritmos más utilizados en AI y XAI.

Algoritmos de AI

Algoritmo de aprendizaje automático	Referencias	Número de artículos
CNN	[31], [39], [43], [46], [53]	5
DNN	[38], [49], [35], [54]	4
Support Vector Machine (SVM)	[31], [38], [50], [32], [34]	5
Random Forest	[31], [38], [48], [50], [32], [55], [41], [45], [57]	9

Tabla 5: Algoritmos de AI más usados en el conjunto de artículos

La Tabla 5 nos indica los algoritmos de AI más usados entre los 27 artículos analizados. En el Anexo 3 se puede visualizar la lista completa de algoritmos de AI usados en el conjunto de artículos.

Algoritmos de XAI

Algoritmo de XAI	Referencia	Número de artículos
LIME	[42], [50], [54], [56]	4
SHAP (SHapley Additive exPlanations)	[39], [42], [33], [41], [52], [49], [56]	7
Deep SHapley Additive Explanations (DeepSHAP)	[38]	1
Grad-CAM++	[46], [31]	2

Tabla 6: Algoritmos de XAI más usados en el conjunto de artículos

La Tabla 6 nos indica los algoritmos de XAI más usados entre los 27 artículos analizados. En el Anexo 3 se puede visualizar la lista completa de algoritmos de XAI usados en el conjunto de artículos.

Naturaleza del conjunto de datos utilizados

El tercer y último enfoque utilizado en la extracción de datos cuantitativos consiste en realizar un análisis de la naturaleza del conjunto de datos que utiliza cada artículo, los posibles escenarios que se tienen es que sea un conjunto de datos existente y publicado para su libre uso o que los datos sean recolectados específicamente para el estudio que se va a realizar (Tabla 7).

	Conjunto de datos existentes y públicos para su uso	Conjunto de datos recolectados para el estudio
Referencia	[37], [38], [39], [42], [55], [46], [44], [32], [50], [48], [51], [33], [41], [34], [45], [35], [54], [57], [36], [47]	[31], [43], [40], [52], [49], [53], [56]
Número de artículos	20	7

Tabla 7: Conjunto de datos clasificados por su naturaleza

Una vez se ha realizado la extracción y clasificación de información se encontró la falta de literatura enfocada en el uso de XAI para el área de Ginecología y Obstetricia. De igual forma, la clasificación de los artículos nos indica una inclinación por los algoritmos más populares de AI (SVM, ANN y RandomForest). Finalmente, para realizar explicaciones se utilizan principalmente dos algoritmos de XAI (LIME y SHAP).

2.2.2 Comprensión de datos

Para nuestra investigación, se utilizó el SisPorto cardiotocograms dataset [44] el cual es un conjunto de datos muy utilizado en el área de obstetricia para la implementación de AI. El data set cuenta con 2126 instancias de cardiotocografías las cuales tienen un total de 22 atributos con una clasificación multivariable que indica la salud fetal “Normal”, “Sospechosa” o “Patológica”.

Descripción del conjunto de datos

Los 22 atributos (Tabla 8) que se obtienen por cada cardiotocografía tienen una descripción enlazada al examen realizado al feto, así como al histograma del resultado de la cardiotocografía.

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, excluyendo cambios periódicos o episódicos, períodos de variabilidad marcada o segmentos de línea base que difieren en más de 25 latidos por minuto. Un feto tiene una línea base normal si su valor se encuentra entre 110 y 160 lpm. Los fetos prematuros tienden a tener valores hacia el extremo superior de este rango y los fetos postérmino hacia el extremo inferior.
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. Reflejan bienestar fetal precedido de un movimiento fetal o contracción.
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 con una duración mayor a 15 segundos.
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%. Se identifica un punto con variabilidad a corto plazo anormal siempre que la diferencia entre dos señales de FCF adyacentes sea inferior a 1 lpm. 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. Estas fluctuaciones se cuantifican visualmente como la amplitud en latidos por minuto.
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. La variabilidad de latido a latido o a corto plazo es 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. Un punto con LTV anormal se identifica cuando la diferencia entre los valores máximo y mínimo de una ventana de 60 segundos no supera los 5 lpm.
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 síntomas de ella.

Tabla 8: Descripción del conjunto de datos [44]

2.2.3 Preparación de datos

Para la preparación de datos se realiza un procedimiento que empieza con la previa visualización de los datos en el cual nos enfocamos en buscar valores nulos o vacíos que a menudo pueden afectar los modelos de AI. Muchos conjuntos de datos suelen tener un desbalance significativo por lo que realizar una visualización de la cantidad de instancias pertenecientes a cada clase es un paso importante que nos ayudará a mejorar la precisión de los modelos de AI que se desean desarrollar.

Limpieza de datos

SisPorto cardiotocograms dataset [58] cuenta con 2126 instancias de tipo float 64, sin valores vacíos o nulos (Figura 10) por lo que no es necesario ninguna acción de eliminación de filas o transformación de datos faltantes los cuales se sugieren realizar en caso de valores NaN.

```

class 'pandas.core.frame.DataFrame'
RangeIndex: 2126 entries, 0 to 2125
Data columns (total 22 columns):
 #   column                Non-Null Count  Dtype
---  ---
 0   baseline_value        2126 non-null   float64
 1   accelerations         2126 non-null   float64
 2   fetal_movement        2126 non-null   float64
 3   uterine_contractions  2126 non-null   float64
 4   light_decelerations   2126 non-null   float64
 5   severe_decelerations  2126 non-null   float64
 6   prolonged_decelerations 2126 non-null   float64
 7   abnormal_short_term_variability 2126 non-null   float64
 8   mean_value_of_short_term_variability 2126 non-null   float64
 9   percentage_of_time_with_abnormal_long_term_variability 2126 non-null   float64
10   mean_value_of_long_term_variability 2126 non-null   float64
11   histogram_width       2126 non-null   float64
12   histogram_min         2126 non-null   float64
13   histogram_max         2126 non-null   float64
14   histogram_number_of_peaks 2126 non-null   float64
15   histogram_number_of_zeroes 2126 non-null   float64
16   histogram_mode        2126 non-null   float64
17   histogram_mean        2126 non-null   float64
18   histogram_median      2126 non-null   float64
19   histogram_variance    2126 non-null   float64
20   histogram_tendency    2126 non-null   float64
21   fetal_health          2126 non-null   float64
dtypes: float64(22)
memory usage: 365.5 KB

```

```

baseline_value        0
accelerations         0
fetal_movement        0
uterine_contractions  0
light_decelerations   0
severe_decelerations  0
prolonged_decelerations 0
abnormal_short_term_variability 0
mean_value_of_short_term_variability 0
percentage_of_time_with_abnormal_long_term_variability 0
mean_value_of_long_term_variability 0
histogram_width       0
histogram_min         0
histogram_max         0
histogram_number_of_peaks 0
histogram_number_of_zeroes 0
histogram_mode        0
histogram_mean        0
histogram_median      0
histogram_variance    0
histogram_tendency    0
fetal_health          0
dtype: int64

```

Figura 10: Detalle del conjunto de datos SisPorto cardiotocograms dataset

Visualización categórica de los datos

En la Figura 11 podemos realizar una visualización de cada categoría de clasificación de los datos, el data set contiene en total 1655 datos de fetos con salud Normal, 295 casos de fetos con salud Sospechosa y 176 casos de fetos con salud Patológica. Evidentemente el

conjunto de datos se encuentra imbalanceado entre sus tres categorías (normal, sospechoso y patológico).

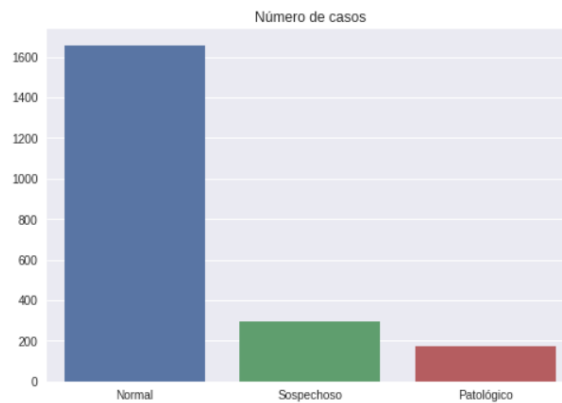


Figura 11: Muestra categórica de la clasificación de la salud fetal en el conjunto de datos

Balanceo de datos

El trabajar con un conjunto de datos imbalanceado implica una pérdida de precisión en los algoritmos de AI [59]. Por lo que un paso para realizar entrenamiento de modelos de AI consiste en realizar un balanceo previo en el conjunto de datos para con esto igualar el número de datos para cada clase.

Existen varios métodos que sirven para realizar un balanceo de datos y con esto obtener un conjunto de datos equilibrado en sus diferentes clases. Algunos de los métodos más utilizados consisten en realizar un balanceo mediante repetición de datos en las clases menos pobladas, eliminación de datos en la clase más poblada y también existen métodos que realizan operaciones matemáticas para repoblar los datos como es el caso del algoritmo SMOTE [60].

Acorde a la necesidad de este estudio de realizar explicaciones de los algoritmos de AI se ha decidido no realizar un balanceo de datos, es importante para nuestro caso de estudio el contar con datos reales que nos indiquen en que medida aportan valor a las decisiones de los algoritmos de AI. Por lo tanto, el repoblar en conjunto de datos con valores sintéticos implica mejorar la precisión de AI, pero reduce la explicabilidad deseada de XAI.

Distribución de los datos

Los histogramas nos ayudan a visualizar la distribución de los datos mediante un conteo de observaciones [61]. En las Figuras 12 y 13 se puede visualizar la distribución de datos que tiene cada variable del conjunto de datos.

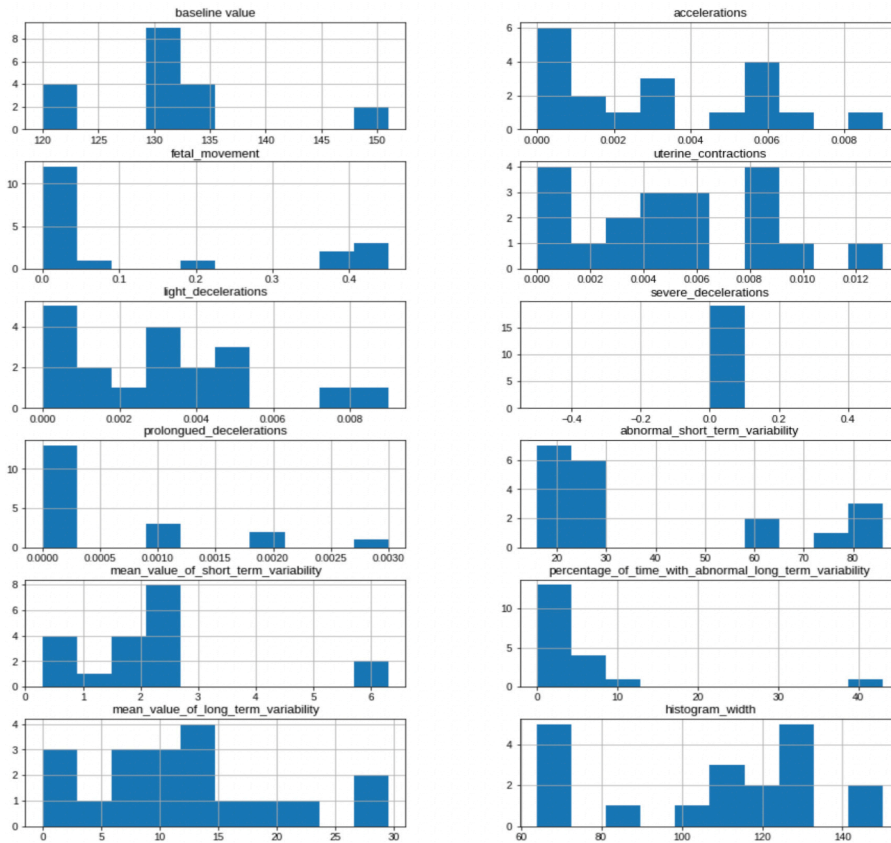


Figura 12: Distribución de datos parte 1

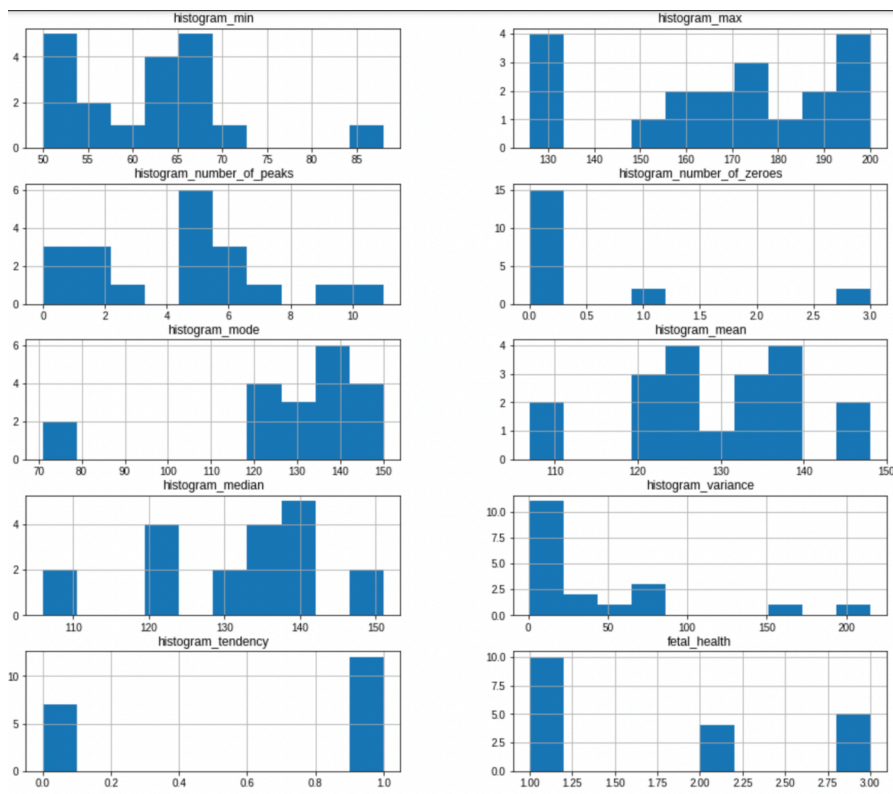


Figura 13: Distribución de datos parte 2

Los histogramas de cada variable indican la existencia de varios valores atípicos que podrían afectar los resultados, debido a esto se puede observar la necesidad de realizar un tratamiento de datos que nos ayuden a disminuir esta dispersión de los datos.

Matriz de correlación

Una matriz de correlación es una forma fácil de resumir las correlaciones entre todas las variables de un conjunto de datos. Para nuestro caso de estudio se ha realizado bajo una escala de colores y números, donde, las correlaciones numéricas que se acerquen a 1 serán de color azul, los valores que no tengan correlación tienen un valor igual a 0 y serán diferenciados por un color crema. Por último, las variables que se correlacionen de manera inversa tendrán un valor negativo que tiende a -1 y su color diferenciador será rojo.

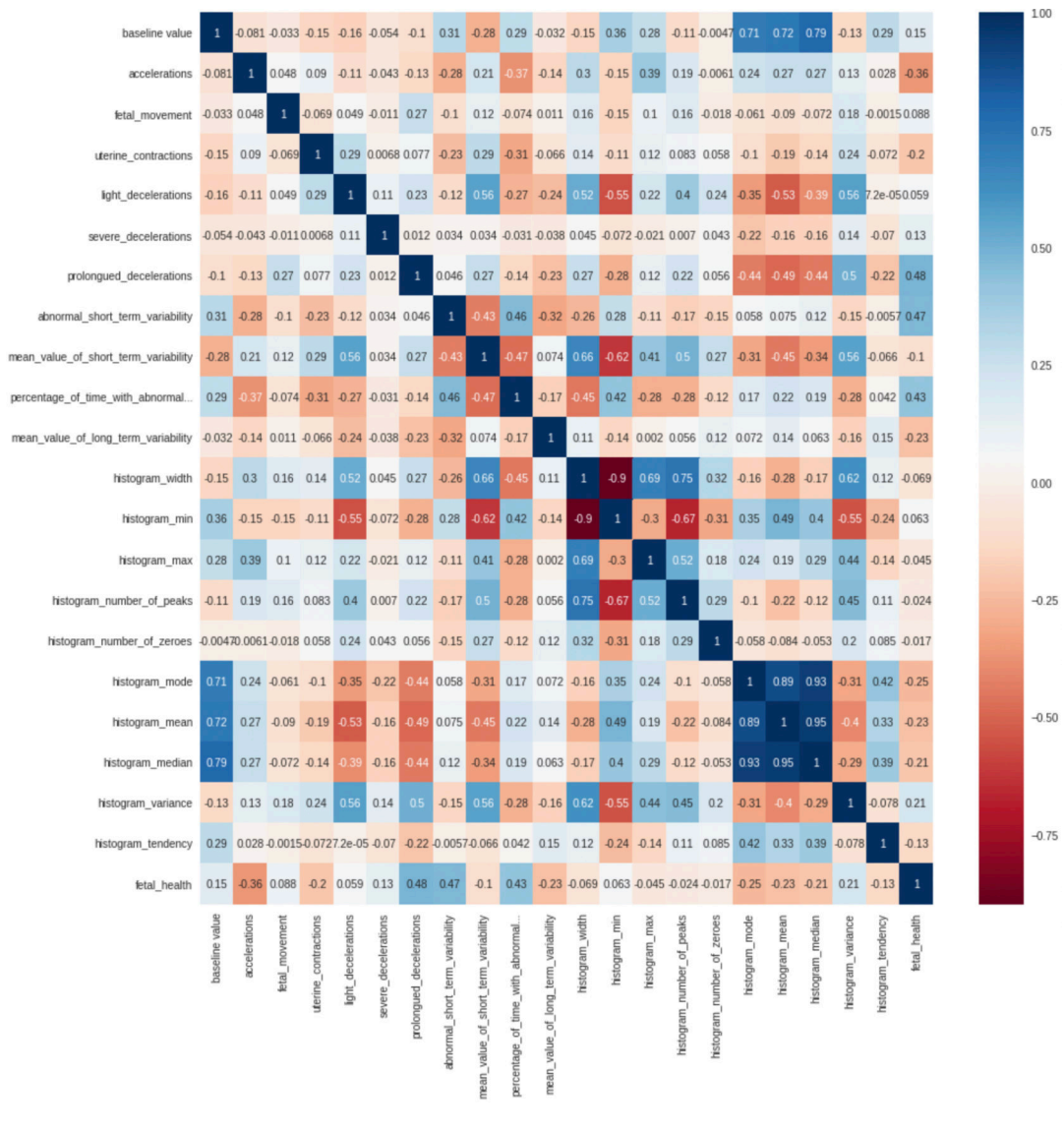


Figura 14: Matriz de correlación del conjunto de datos

Los valores de media, moda y mediana están altamente correlacionados, esto tiene mucho sentido ya que son valores que se obtienen de un histograma. El baseline value de igual forma tiene una correlación con datos del histograma como son la moda media y varianza. Por otro lado, se puede notar una relación inversa en varias variables y muchos otros valores que tienden a una correlación muy baja.

Fuera de los datos propios del histograma, una correlación notable la podemos encontrar entre valor medio de la variabilidad a corto plazo (`mean_value_of_short_term_variability`) con las desaceleraciones ligeras (`light_decelerations`) y el valor mínimo del histograma.

Las variables que pertenecen al histograma muestran la mayor parte de las correlaciones entre si, los valores obtenidos de la cardiocografía muestran correlaciones positivas y negativas muy bajas entre si, sin embargo, cuentan con correlaciones un poco más altas con valores del histograma.

Preparación de datos

Selección de características

La selección de características es un factor muy influyente en el aprendizaje automático. Para que los algoritmos funcionen correctamente y proporcionen predicciones casi perfectas, es decir, para mejorar el rendimiento de un modelo predictivo, se requiere la selección de características [45]. Se deben eliminar las características irrelevantes o redundantes (Figura 13). Hasta un cierto número de características, la precisión de un clasificador aumenta, pero hay un umbral a partir del cual comienza a disminuir. El uso de demasiadas o muy pocas características puede conducir al problema de alta varianza y alto sesgo lo que conduce a tener problemas de sobreajuste o subajuste (Figura 15). Por lo tanto, encontrar el mejor subconjunto de características es un paso importante en la preparación de los datos.

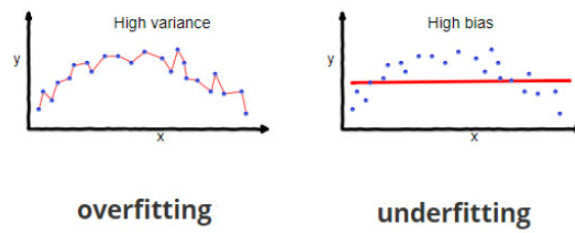


Figura 15: Ejemplo gráfico de alta varianza y alto sesgo. Imagen obtenida de: towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

En esta sección buscaremos eliminar características irrelevantes del conjunto de datos (Figura 16) con el fin de evitar la complejidad innecesaria en el entrenamiento del modelo de aprendizaje de máquina. Existen varios métodos que nos ayudan a la selección de características los cuales se dividen en supervisados y no supervisados [46].

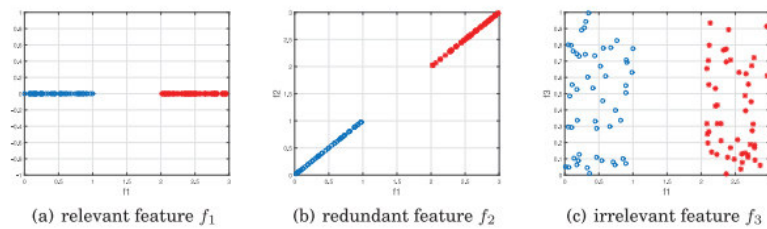


Figura 16 Ejemplo gráfico de características relevantes, redundantes e irrelevantes [46]

Cuando se realiza una selección de características basada en estadísticas, es importante seleccionar el método que se va a utilizar en función de los tipos de datos de las variables de entrada y salida. En la Figura 17 se muestra un árbol de decisión para ayudar en la selección del método adecuado acorde a las entradas y las salidas:

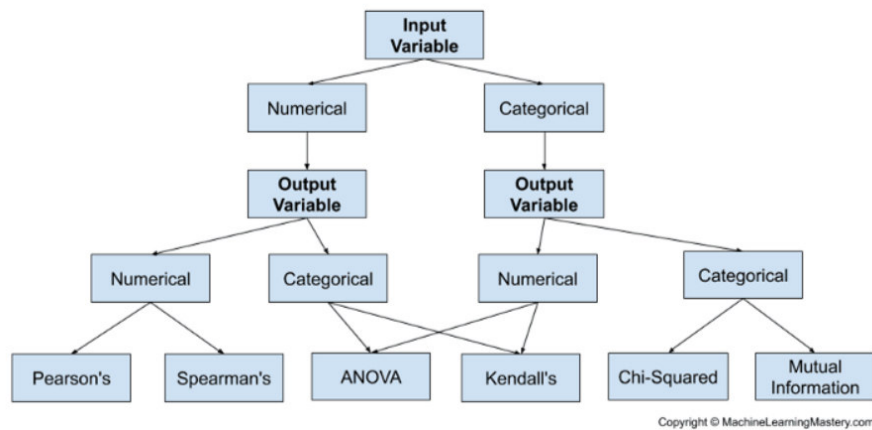


Figura 17: Árbol de decisión para escoger el método de selección de categoría, Imagen obtenida de MachineLearningMastery.com

ANOVA

El análisis de varianza es un método estadístico que prueba si las diferentes categorías de entrada tienen valores significativamente diferentes para la variable de salida [47]. ANOVA permite el análisis de múltiples grupos de datos para determinar la variabilidad entre muestras y dentro de muestras, con el fin de obtener información sobre la relación entre las variables dependientes e independientes.

SelectKBest es una clase de la librería Scikit-learn que nos ayuda a determinar las mejores características de un conjunto de datos determinados [48]. El método SelectKBest selecciona las características de acuerdo con la puntuación k más alta. La selección de las mejores características es un proceso importante cuando preparamos un gran conjunto de datos para el entrenamiento, esto nos ayuda a eliminar los atributos menos importantes del conjunto de datos y con esto obtener próximas ventajas como una reducción en los tiempos de ejecución de algoritmos de AI.

Debido al caso de este estudio se ha decidido realizar una selección de características únicamente de los atributos del histograma, esto debido a que los valores obtenidos de la cardiocografía son esenciales para realizar explicaciones de los algoritmos de AI y así conocer los atributos más influyentes para determinar la salud fetal.

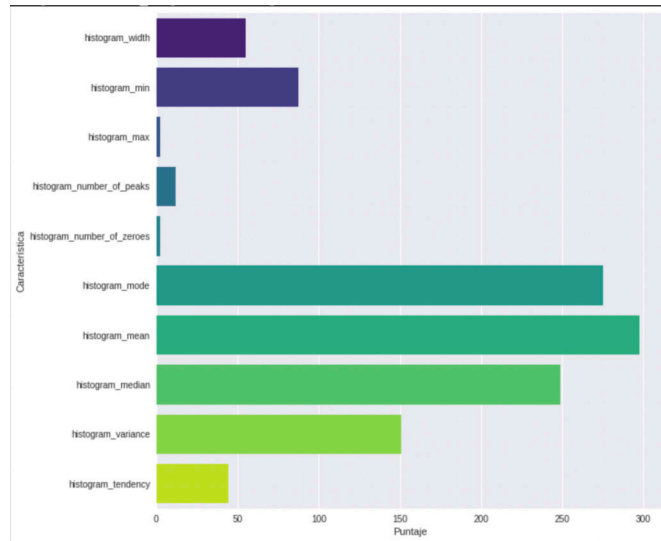


Figura 18: Selección de características del histograma usando el algoritmo SelectKBest

En la Figura 18 se puede observar que los atributos como: porte, mínimo, máximo, número de picos, número de ceros y tendencia del histograma tienen una puntuación k muy baja por lo que implica su poca relevancia en el conjunto de datos y son seleccionables para descartarlos y mantener únicamente los datos más importantes del histograma y los datos propios de la cardiocografía.

Estandarización de características

Durante la visualización de datos notamos la presencia de valores atípicos en conjunto de datos, esto podría ser un factor muy influyente al momento de realizar el entrenamiento de un modelo de AI, debido a esto surge la necesidad de realizar una estandarización en el conjunto de datos. En la Figura 19 podemos visualizar la distribución distante que tienen las características del conjunto de datos.

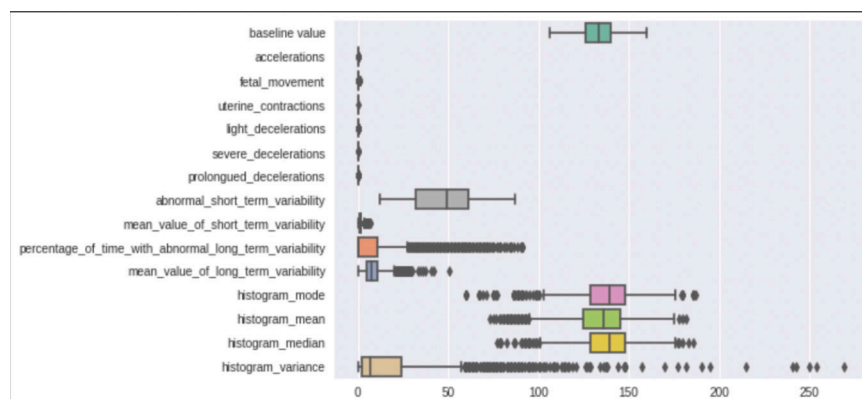


Figura 19: Bloxpot que indica la varianza en la distribución de datos de cada atributo

La estandarización de características es el proceso de adecuar los valores de un conjunto de datos a una escala en la cual los datos se encuentren parcialmente juntos. Este proceso consiste en la reducción de valores atípicos o valores con una distribución muy distante. Existen varios métodos de estandarización, algunos de ellos consisten en normalizar los valores entre una escala de -1 y 1 con una distribución normal y otros que realizan una eliminación de su media y escala a una varianza unitaria como el Standard Scaler de la librería Sklearn [48]

Los modelos de AI necesitan una estandarización de datos para mejorar su precisión y con esto obtener resultados más eficientes en sus predicciones. Sin embargo, los algoritmos de AI basado en arboles de decisión cuentan con sus propios métodos de estandarización [66], por lo tanto, se realiza una estandarización de características para los algoritmos de AI que no estén basados en arboles de decisión.

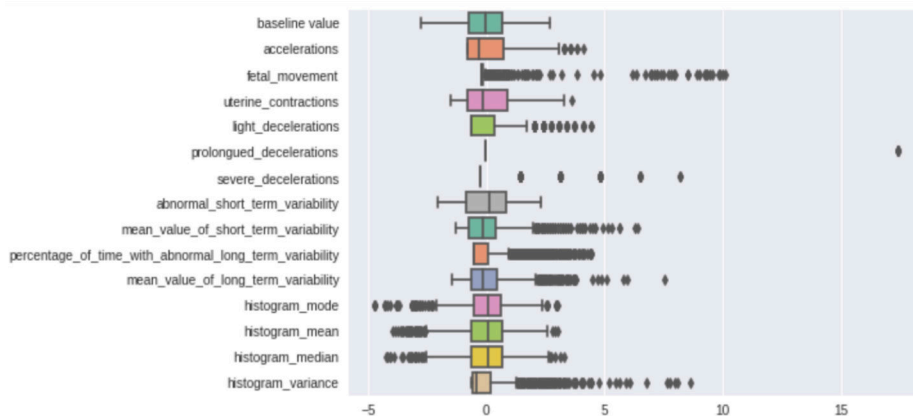


Figura 20: Boxplot que indica la varianza en la distribución de los datos de cada atributo luego de realizar una estandarización mediante el algoritmo StandarScaler de la librería sklearn

En la Figura 20 se muestra una distribución de datos luego de realizarse una estandarización. Se puede observar que se encuentra más equilibrada con respecto a los datos sin estandarización (Figura 19) y con esto podemos resolver que el conjunto de datos se puede aprovechar de mejor forma por los algoritmos de AI que no estén basado en arboles de decisión.

2.2.4 Modelado

En esta fase del modelo CRISP DM se debe utilizar toda la información obtenida con anterioridad. Con esto, llevamos a cabo la idea del planteamiento del negocio donde nos enfocamos en utilizar el algoritmo SHAP de XAI para realizar explicaciones sobre un algoritmo de AI el cual será seleccionado acorde a una comparación de los resultados de diferentes que serán seleccionados.

Algoritmo de AI

Para la selección de los algoritmos de AI se excluyen aquellos que son interpretables por naturaleza, en este aspecto se consideran clasificadores lineales o modelos interpretables basados en arboles de decisión. Esta exclusión se realiza con el fin de cumplir con los requerimientos del modelo SHAP, el cual es un modelo Post-Hoc de XAI que utiliza como entrada un modelo de AI que expresa su procedimiento como una caja negra.

Acorde a la SLR previamente realizada se observa la concurrencia de uso de metodologías tradicionales de AI para realizar explicaciones con el uso de algoritmos post-hoc de XAI, basado en esto, se ha decidido seleccionar dos algoritmos de AI tomando en cuenta los más usados por otros autores en trabajos relacionados siempre que se traten de algoritmos supervisados.

El primer algoritmo de AI utilizado fue Support Vector Machine (SVM) el cual ha sido utilizado en 5 trabajos relacionados (Tabla 5). SVM se trata de un conjunto de métodos de aprendizaje supervisado que son, en gran cantidad, utilizados para realizar clasificaciones, regresiones y detección de valores atípicos [48]. Funciona de manera correcta y es eficiente con conjuntos de datos multiclase.

El segundo algoritmo seleccionado es Random Forest que ha sido utilizado por 9 autores en diferentes trabajos relacionados (Tabla 5). El algoritmo de Random Forest se trata un modelo basado en arboles de decisión que realiza un ajuste de una serie de clasificadores de arboles para submuestrear el conjunto de datos y obtener el promedio de estos, con lo cual realiza una mejora predictiva y controla lo que conocemos como sobreajuste [48].

El ultimo algoritmo seleccionado fue redes neuronales las cuales son altamente utilizada en el modelado de algoritmos de AI, por lo cual, es un algoritmo que ha sido considerable indispensable en nuestra investigación. Las redes neuronales son parte del aprendizaje

supervisado, actualmente se les considera la subárea más relevante del aprendizaje automático. Son responsables del ajetreo y el bullicio actual de la AI porque pueden evaluar grandes cantidades de datos no estructurados particularmente bien y tiene la capacidad de encontrar patrones en los mismos [67].

En nuestro caso de estudio se puede analizar una entrada de los 15 atributos seleccionados con una salida de clasificación multiclase que representan la salud fetal (Normal, Sospechosa, Patológica) el cual mediante diferentes capas neuronales se encargan del reconocimiento de patrones y tomar la decisión de clasificación. Para una visualización del proceso de ANN en nuestro caso de estudio se puede analizar la Figura 21

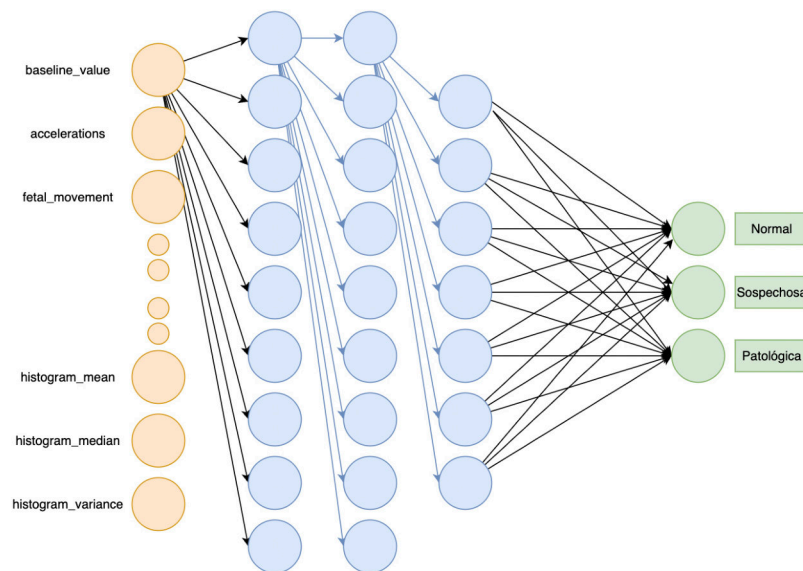


Figura 21: Red Neuronal Artificial para evaluar la salud fetal

Para realizar una selección de un solo algoritmo entre los 3 candidatos se usó un plan de pruebas en el cual se evaluará el algoritmo con mejor rendimiento para el conjunto de datos. El plan de pruebas consistió en realizar una comparación de las distintas métricas de evaluación de modelos de AI mediante una selección de las métricas adecuadas.

Algoritmo de XAI

SHAP es un algoritmo de XAI que se ejecuta con la idea post-hoc, es decir, se usa con la salida de un algoritmo de caja negra de AI. Para esto se realiza un plan de prueba que nos permita para evaluar los algoritmos de AI y obtener el algoritmo que cuenta con los resultados más favorables.

Plan de pruebas para la evaluación de algoritmos de AI

Para realizar una evaluación de los diferentes algoritmos de AI se hace uso de distintas métricas para medir el rendimiento del modelo y obtener valores numéricos que nos permitan realizar una comparación con los otros modelos de aprendizaje automático. Para realizar una selección de las métricas que se usarán se considera el conjunto de datos y las métricas que son aplicables para el mismo.

Matriz de confusión

Una de las métricas más conocidas para evaluar un algoritmo de AI es realizar una matriz de confusión la cual es una Tabla que mide el rendimiento del modelo mostrando las posibles fallas y dando pistas de donde puede el modelo confundir los resultados. Está compuesta por 4 partes que se debe conocer para realizar una correcta interpretación de la matriz de confusión [68]:

True Positive: Los valores verdaderos positivos (TP, por sus siglas en inglés) corresponde a los valores positivos que fueron predichos correctamente por el modelo de AI.

True Negative: Los valores verdaderos negativos (TN, por sus siglas en inglés) corresponde a los valores negativos que fueron predichos correctamente por el modelo de AI.

False Positive: Los valores falsos positivos (FP, por sus siglas en inglés) corresponde a los valores positivos que fueron predichos como negativos.

False Negative: Los valores falsos negativos (FN, por sus siglas en inglés) corresponde a los valores negativos que fueron predichos como positivos.

		VALORES REALES	
		Positivo	Negativo
VALORES PREDICHOS	Positivo	TP	FP
	Negativo	FN	TN

Figura 22: Matriz de confusión

Precisión

Esta métrica es comúnmente una de las más usadas para evaluar el rendimiento o calidad de un modelo de AI [68]. Nos indica que tan precisos son los resultados obtenidos en una escala de 0 a 1 donde, 0 significa que el modelo no es confiable y mientras más alto sea su valor será más confiable. La precisión de un modelo se obtiene mediante la división de los valores TP sobre la suma de los valores TP más los valores de FP (Ecuación 1).

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Ecuación 1: Formular para calcular la precisión de un modelo de AI

Recall

El recall nos ayuda a medir la capacidad que tiene un modelo de AI para realizar una detección de muestras positivas. Se encarga de cuantificar las predicciones que fueron realizadas correctamente de todas las predicciones positivas que podrían haberse realizado. Para obtener el Recall de un modelo se debe dividir los valores TP sobre la suma de los valores TP más los valores FN (Ecuación 2).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Ecuación 2: Formula para calcular el recall de un modelo de AI

F1-score

La métrica F1-Score es muy utilizada para la evaluación de un modelo de AI ya que permite obtener un resumen del rendimiento del modelo en general, genera una forma de unificar las propiedades de la precisión y el recall del modelo.

$$F1 = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

Ecuación 3: formula para el calculo del valor f1-score de un modelo de AI

2.2.5 Despliegue

Para el despliegue se propone el planteamiento de una estrategia de implementación y control que se regiría en un proceso definido que permita realizar una evaluación de la salud fetal. Este proceso tiene inicio con la introducción de los datos en el sistema hasta el monitoreo del funcionamiento de los modelos de AI y XAI, para lo cual se realizaría un sistema que cumpla con la automatización de este flujo.

El sistema recibiría una entrada en formato CSV que, en su contenido, incluiría una instancia de datos de la salud de un feto los cuales se obtienen mediante la salida de la cardiotocografía, el sistema leería estos datos y usaría el modelo de AI que haya sido seleccionado para su implementación y se encargaría de la predicción de la salud fetal.

Una vez el médico especialista obtenga un resultado de la salud fetal el cual puede ser Normal, Sospechoso o Patológico debería ser capaz de entender el ¿por qué? De la decisión por parte del algoritmo de AI. Para esto se plantea la utilización de unas técnicas de visualización que nos permitan indicar los resultados obtenidos por XAI indicando de manera progresiva las alteraciones que tiene la salud fetal con respecto a cada variable y como se logra obtener ese resultado.

Como parte del mejoramiento del sistema, se obtendría comentarios de los usuarios médicos especialistas que usen el software, esto con el fin de entender si existen problemas de predicción y poder mejorar el algoritmo de AI o, caso contrario, si se encuentran problemas de interpretación, realizar ajustes a la visualización de las explicaciones que nos ofrece el algoritmo SHAP de XAI.

3 PRUEBAS, RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Pruebas

En esta sección se plantea la sección de pruebas con un médico especialista en el área de obstetricia para lo cual se ejecuta el Modelo de Aceptación Tecnológico (TAM, por sus siglas en inglés) [49] el cual nos ayuda a obtener una retroalimentación de la ejecución de los algoritmos de AI y XAI.

TAM tiene como objetivo obtener comentarios de los usuarios que hacen uso de una herramienta tecnológica [49]. Con este fin plantea la idea de recolectar información sobre la facilidad de uso y la utilidad percibidas lo cual engloba las percepciones que tienen los usuarios al usar una herramienta tecnológica.

Facilidad de uso percibida

¿Qué tan fácil de entender le parecieron las explicaciones obtenidas?

¿Con qué facilidad podría implementar estas explicaciones en su trabajo?

Utilidad percibida

¿Qué tan útiles le parecieron las explicaciones obtenidas?

¿Qué tan útil considera que sería implementar en su trabajo las explicaciones obtenidas?

3.2 Resultados

3.2.1 Modelos de AI

Acorde a las métricas planteadas para para determinar el rendimiento de los algoritmos de AI a usarse, se graficó una matriz de confusión de cada modelo con el cual podamos ver sus porcentajes de precisión, así como el porcentaje en el que el modelo tiende a fallar. Las matrices que se indican en las Figuras 23, 24, 25 pertenecen a los algoritmos de aprendizaje automático previamente seleccionados SVM, Random Forest, ANN respectivamente.

SVM

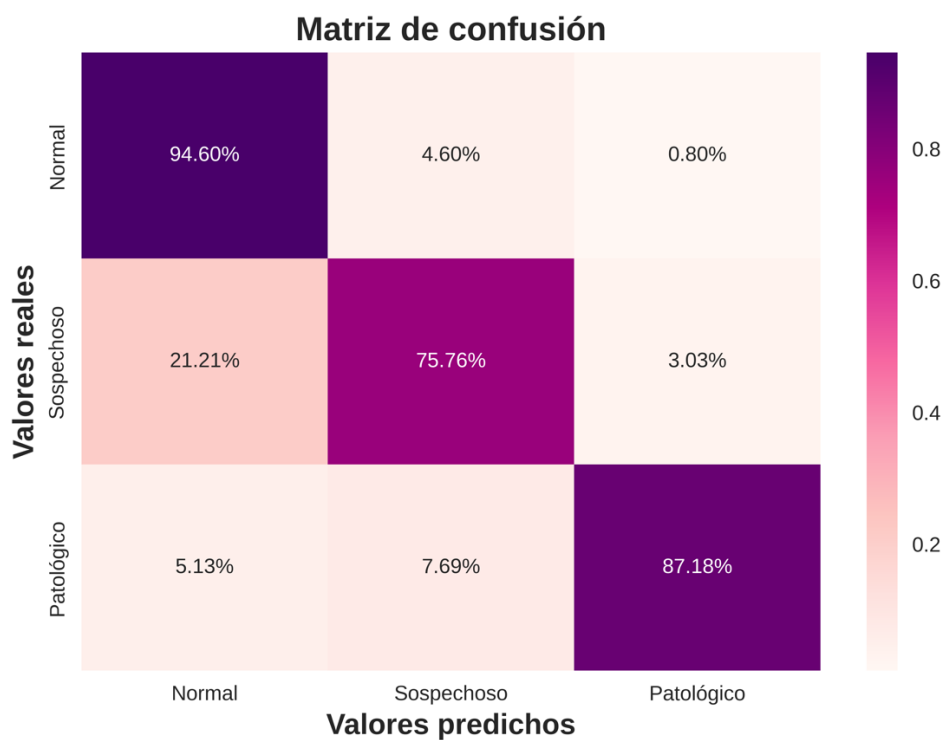


Figura 23: Matriz de confusión del modelo SVM

Radom Forest (RF)

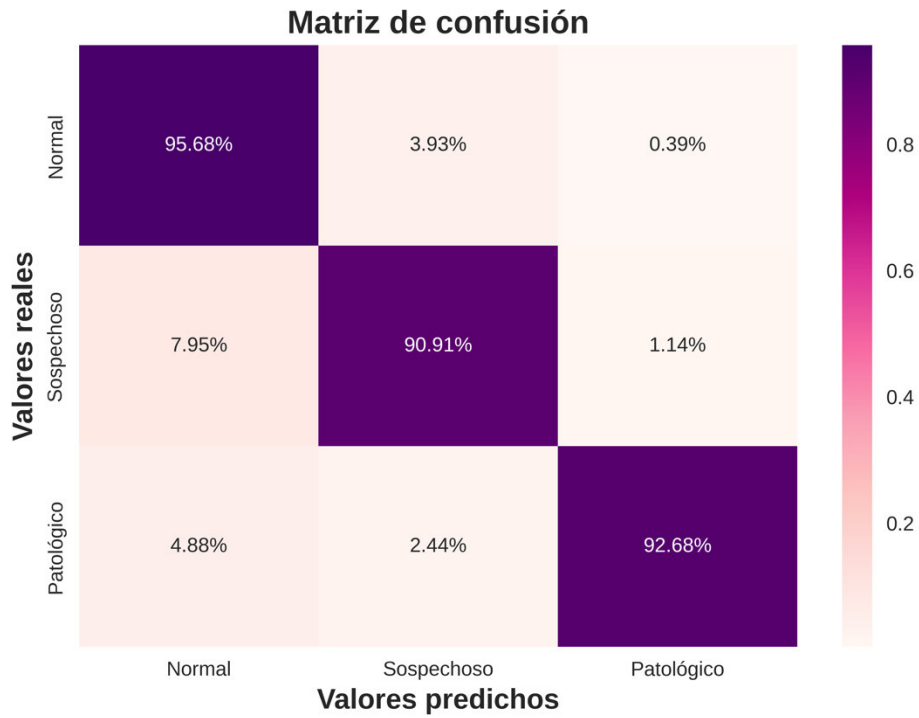


Figura 24: Matriz de confusión del modelo RF

Red Neuronal Artificial (ANN)

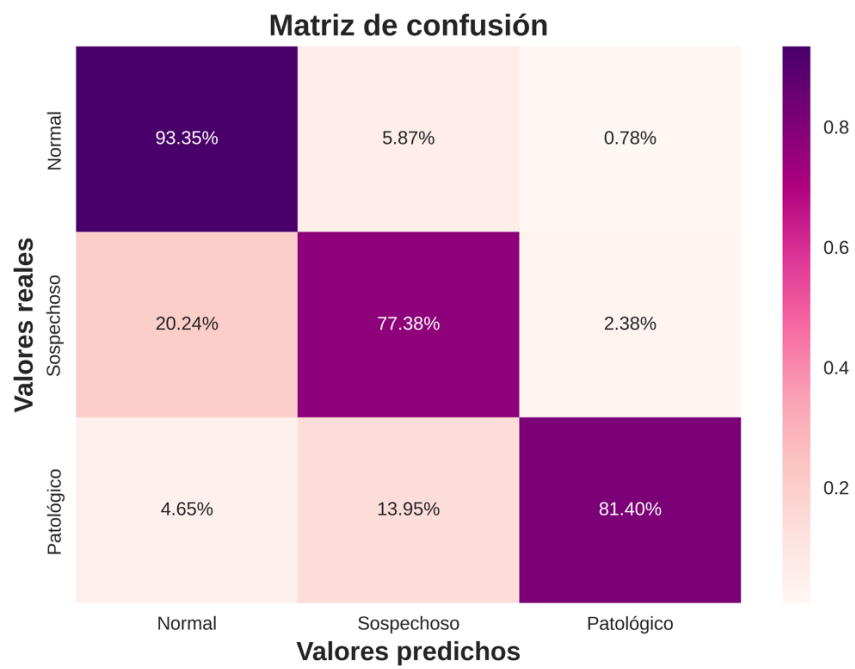


Figura 25: Matriz de confusión del modelo ANN

Las métricas de Precision, Recall y F1-Score se usan para plantear una comparación del rendimiento entre cada uno de los algoritmos y con esto evaluar el algoritmo con mejores resultados de ejecución.

	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	0,71	0,79	0,95	0,69	0,76	0,94	0,7	0,77

Tabla 9: Resultados de la ejecución de los diferentes algoritmos de AI con las clases Normal (N), Sospechoso (S) y Patológico (P).

La Tabla 9 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 SHAP de XAI que se implementará a continuación.

Para realizar una selección de las métricas más influyentes en nuestro caso de estudio fue importante enfocarse en el conjunto de datos utilizado ya que existen métricas que no tienen una confianza relevante si el conjunto de datos está desbalanceado, tal es el caso de la precisión [59]. Por esta razón, el Recall de cada modelo fue la métrica comparación seleccionada.

	Recall		
	N	S	P
SVM	0,95	0,74	0,83
RF	0,98	0,79	0,93
ANN	0,95	0,69	0,76

Tabla 10: Comparación Recall de los modelos de AI.

Para la clase Normal el modelo de RF tiene una ventaja sobre los otros dos modelos que están iguales con su Recall, En la clase Sospechosa, una vez más, el modelo RF tiene una ventaja sobre los otros dos modelos, donde SVM le sigue y en último lugar ANN. Finalmente, para evaluar el Recall de la clase Patológica se puede observar que el Recall de RF es también superior entre los 3 modelos. Razón por la cual se selecciona a RF como el algoritmo con mejores resultados en nuestro caso de estudio.

3.2.2 Modelo de XAI

El modelo de RF que ha sido seleccionado como modelo más eficiente de AI tiene una salida de clasificación de caja negra; por lo que, para conocer el motivo que inclinó al modelo de AI a tomar esa decisión se necesita de un modelo Post-hoc de XAI. En nuestro caso particular, como se ha realizado el planteamiento de este trabajo, se utilizó el algoritmo SHAP de XAI que cumple con este requisito de evaluación.

SHAP plantea la posibilidad de realizar explicaciones locales para diferentes tipos de modelos [24], la selección del Explicador se ha realizado con la teoría de que cualquier algoritmo de AI pudo tener el mejor rendimiento y no exclusivamente un algoritmo basado en arboles de decisión. Por lo que, se descarta el uso de un Explicador de árbol (TreeExplainer). En este caso, se realiza la ejecución de un KernelExplainer el cual es agnóstico al modelo.

A pesar del coste computacional que implica la ejecución de un KernelExplainer, se tiene la ventaja de realizar una explicación local por lo que se escoge únicamente 1 instancia por cada clase de clasificación (Normal, Sospechosa, Patológica). Las instancias han sido seleccionadas del conjunto de datos de prueba, siendo:

- La instancia 400 la correspondiente a la clase Normal,
- La instancia 421 la correspondiente a la clase Sospechosa
- La instancia 25 correspondiente a la clase Patológica.

Instancia Normal

La Tabla 11 muestra los atributos con sus respectivos valores de una instancia que representan la clase Normal.

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

Tabla 11: Instancia de la clase Normal

Instancia Sospechosa

La Tabla 12 muestra los atributos con sus respectivos valores de una instancia que representan la clase Sospechosa.

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

Tabla 12: Instancia de la clase Sospechosa

Instancia Patológica

La Tabla 13 muestra los atributos con sus respectivos valores de una instancia que representan la clase Patológica.

1. baseline_value	133.0	9. mean_value_of_short_term_variability	0.3
2. accelerations	0.000	10. percentage_of_time_with_abnormal_long_term_variability	84.0
3. fetal_movement	0.000	11. mean_value_of_long_term_variability	3.5
4. uterine_contractions	0.000	12. histogram_mode	134.0
5. light_decelerations	0.000	13. histogram_mean	134.0
6. prolonged_decelerations	0.000	14. histogram_median	135.0
7. severe_decelerations	0.000	15. histogram_variance	0.0
8. abnormal_short_term_variability	73.0		

Tabla 13: Instancia de la clase Patológica

Explicaciones de cada instancia

Un aspecto fundamental para entender las explicaciones que realiza el algoritmo SHAP de XAI es obtener la media del valor de cada atributo del conjunto de datos en general. Estos valores nos sirven para realizar una comparación con los valores de cada instancia y determinar si el aumento o disminución de ese valor implica que el atributo tenga algún factor determinante en el cambio de clase.

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

Tabla 14: Valores medios de cada atributo en el conjunto de datos

Explicaciones

Explicaciones con el algoritmo SHAP correspondientes a una clase Normal.

El modelo SHAP cuenta con una gran cantidad de posibilidades gráficas para realizar explicaciones. Para explicaciones locales es común utilizar graficas de decisión, las cuales nos indican como cambia la curvatura de la gráfica acorde al valor de cada atributo hasta llegar su pico a la clase predicha por el algoritmo de AI.

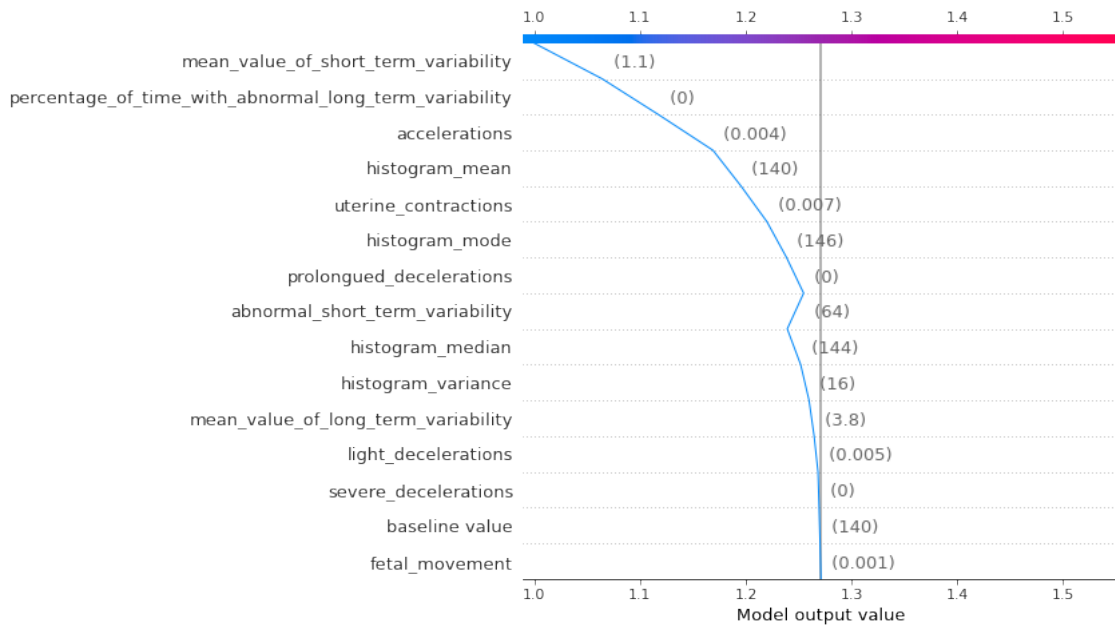


Figura 26: Gráfico de decisión para realizar explicaciones de la instancia Normal

Para realizar una interpretación de la Figura 26 podemos analizar como cambia la curvatura del grafico empezando por una clase definida por la media de las clases obtenida de todo el conjunto de datos y la cual es 1.27, representando $E[f(x)]$ (Figura 27). Con esto podemos determinar que el punto de partida de un examen médico corresponde a la salud Normal de un feto y de aquí en adelante la curvatura cambia acorde al valor de cada atributo.

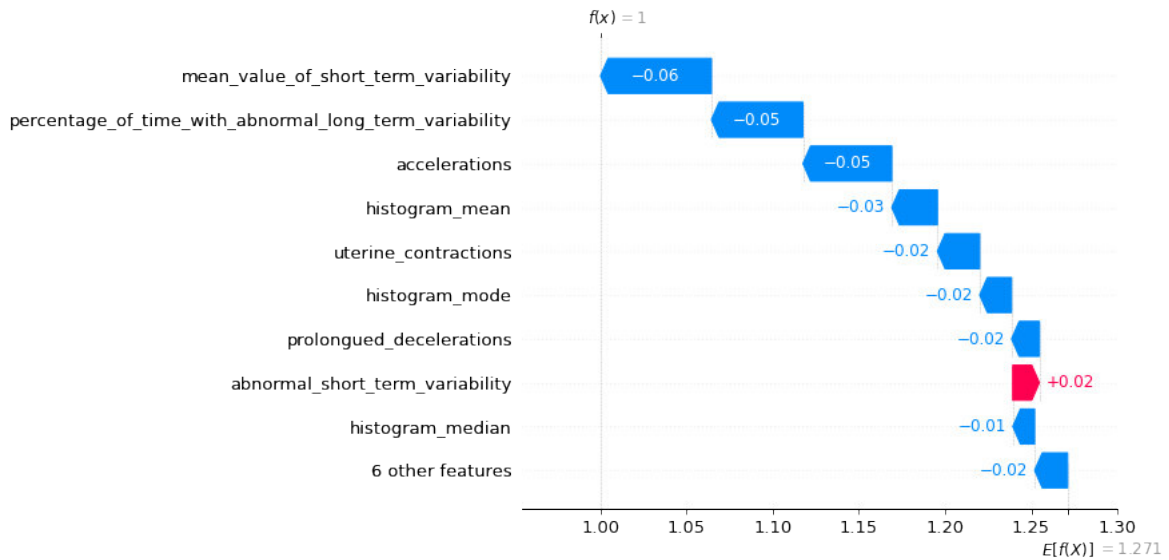


Figura 27: Gráfico de cascada para realizar explicaciones de la instancia Normal

En la Figura 27 se expresa con porcentajes logarítmicos la influencia de cada variable. Utilizando la grafica de decisión podemos realizar una comparación de cada atributo de la Tabla 14 para determinar si el valor es mayor o menor y con esto entender si influye de

manera positiva para una salud normal (Color azul) o influye de manera negativa a la salud normal del feto (Color rojo) y tiende a ser una salud sospechosa o patológica.

Para esta primera instancia se realiza una explicación en la cual, empezando por el valor de 1.27, la salud del feto termina siendo positiva en lo que llamamos una salud Normal que pertenece a la clase 1 en la gráfica. Únicamente podemos visualizar que las observaciones de cada atributo obtenidas en esta cardiotocografía influyen de manera positiva a la salud normal del feto, con la excepción de los términos cortos de variabilidad anormal.

Explicaciones con el algoritmo SHAP correspondientes a una clase Sospechosa.

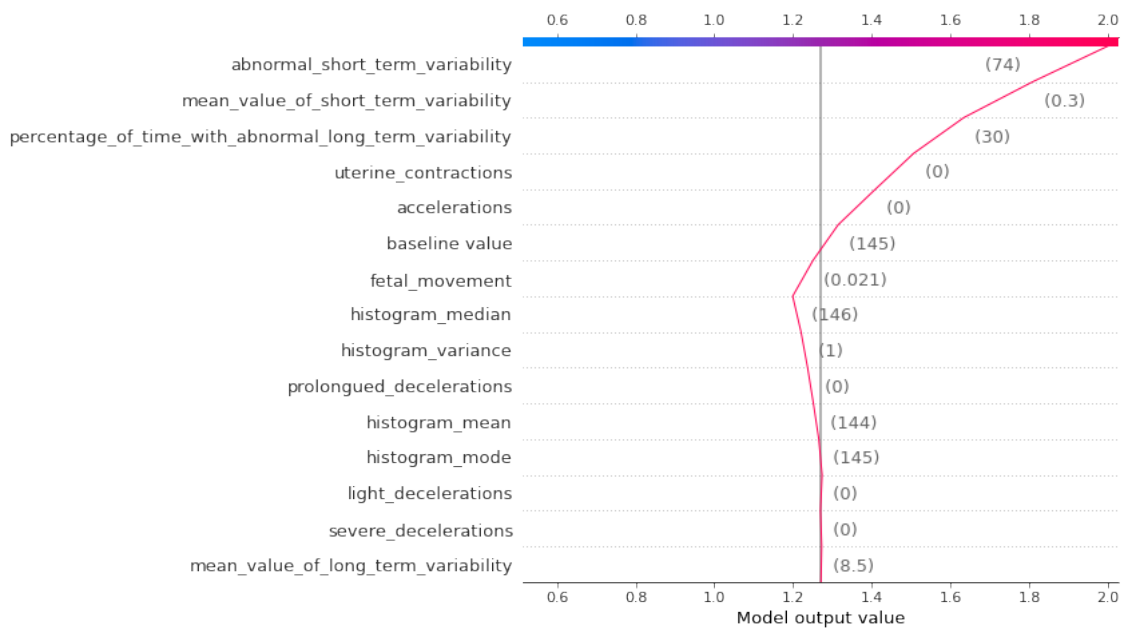


Figura 28: Gráfico de decisión para realizar explicaciones de la instancia Sospechosa

Para la instancia sospechosa se puede evidenciar como tiene la curva a cambiar acorde al valor de los atributos, partiendo desde un valor que se puede considerar normal a la salud fetal hasta llegar a sospechosa siendo los términos de variabilidad anormal el factor más influyente como podemos observar en la Figura 28 y Figura 29.

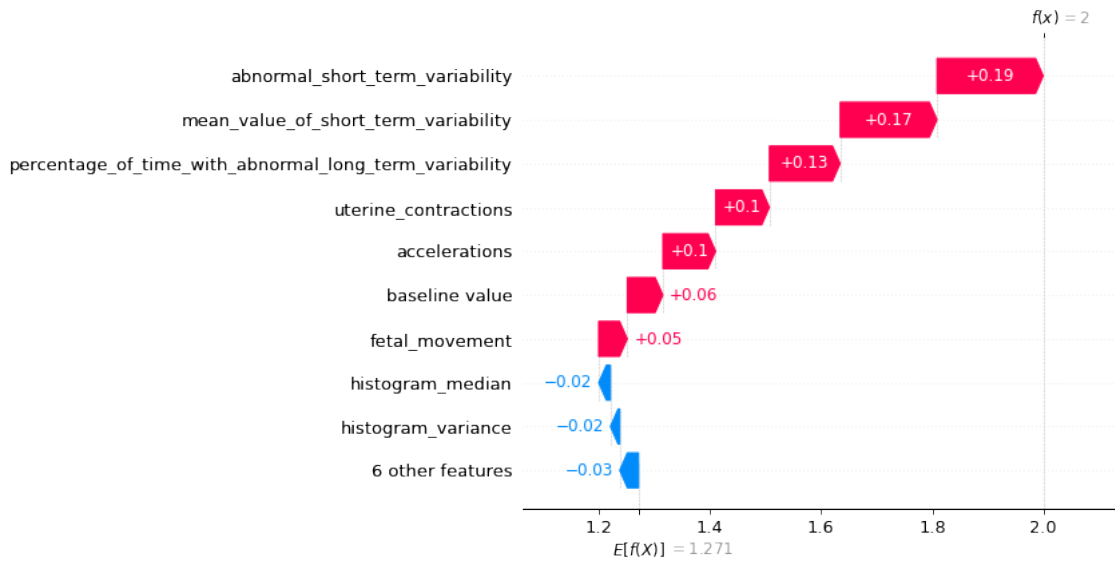


Figura 29: Gráfico de cascada para realizar explicaciones de la instancia Sospechosa

Explicaciones con el algoritmo SHAP correspondientes a una clase Patológica.

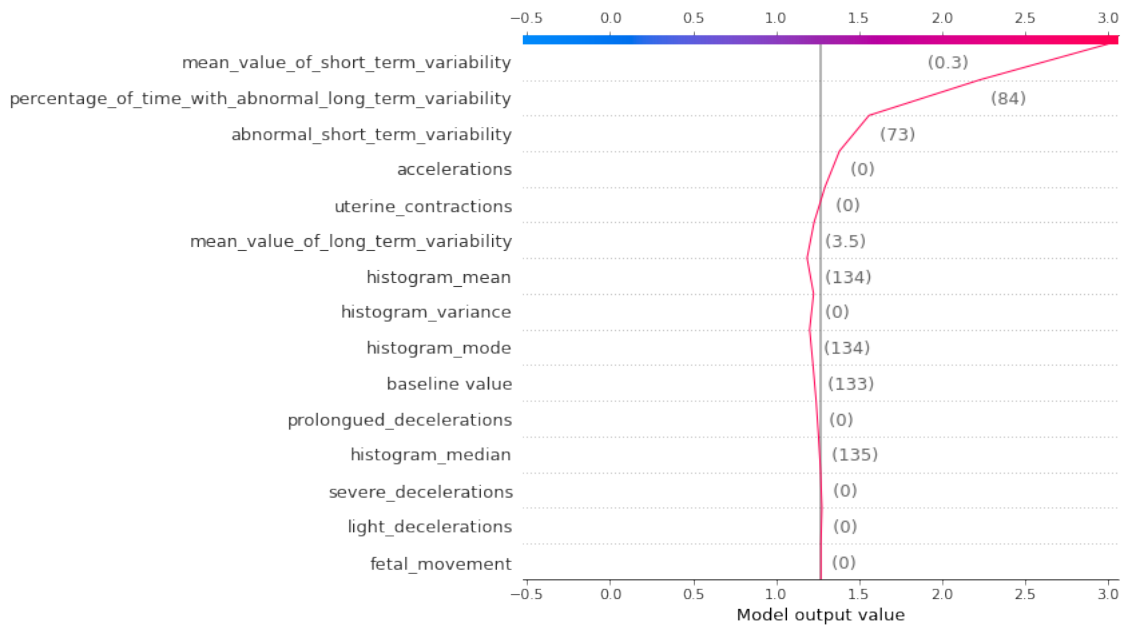


Figura 30: Gráfico de decisión para realizar explicaciones de la instancia Patológica

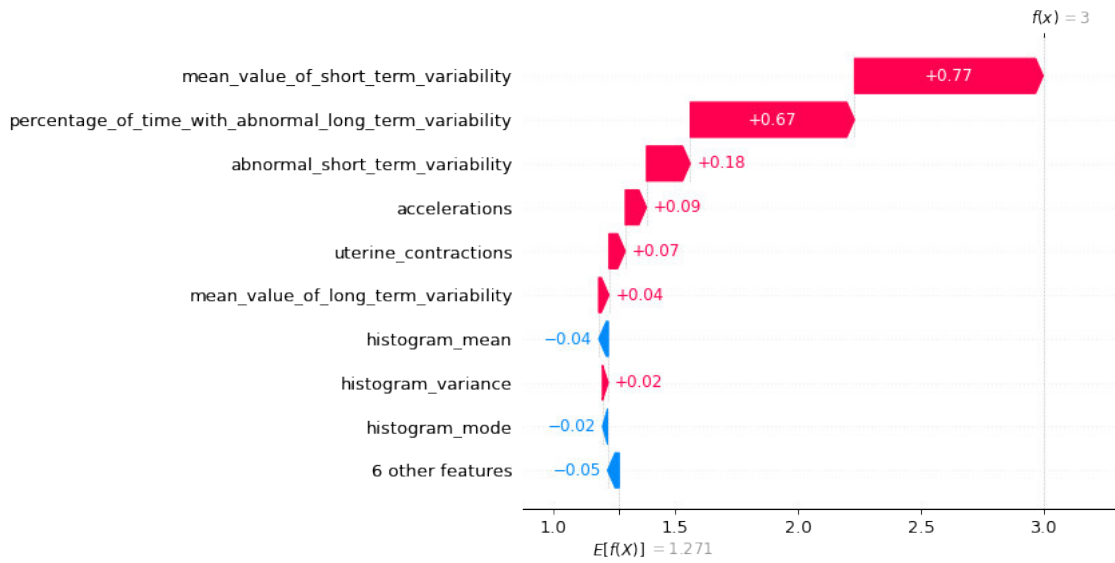


Figura 31: Gráfico de cascada para realizar explicaciones de la instancia Patológica

Finalmente podemos observar como se realiza la clasificación para una instancia patológica, nuevamente parte desde un valor de la salud normal y nos indica la gráfica como tiende a cambiar la clasificación de la salud acorde a los valores de cada atributo, en este caso, como indican las Figuras 30 y 31 podemos observar la variación de la curva mostrando que atributos influyen en la decisión del algoritmo e indicando una jerarquía de los más influyentes para determinar una salud patológica.

3.2.3 Evaluación médica

Para validar la capacidad explicativa que tiene el algoritmo SHAP de XAI se realizó la evaluación de los resultados con una médica con especialidad en Ginecología y Obstetricia (Formación en Patología del tracto genital inferior. Maestría en climaterio y menopausia. Autora de u Caso Clínico).

Opinión médica del método SHAP de XAI

“Bueno, si bien es cierto no es una comparación, pero el método tiene un poco más de claridad en cuanto al observar una curva es de gran ayuda visualizar como influye lo que es normal y como se dirige hacia lo que es patológico. Evidentemente en la gráfica se comprende bastante bien, se ve el desarrollo que es bastante útil. Me gusta también y pienso que es un método que nos puede servir mucho para la toma de decisiones. Sobre todo en esta última parte en la que recalcamos que la persistencia, por ejemplo, de las desaceleraciones donde se marca un tiempo más prolongado y está al final de la curva y eso permite que tomemos decisiones correctas en el momento adecuado. Así que es un método a mí parecer muy bueno y que el permitirnos comparar nos da una curva de seguimiento como para poder tomar desde la base de lo normal y poder ver la evolución de lo que estamos analizando.”

Preguntas acordes a la metodología TAM

Facilidad de uso percibida

¿Qué tan fácil de entender le parecieron las explicaciones obtenidas?

“Bastante claras y fáciles, repito el gráfico ayuda y aporta un montón para tener claro que es lo que se está analizando, qué se va a revisar y en qué punto de la curva nos encontramos para poder revisar los resultados que se obtienen en las diferentes comparaciones de los estudios de cardiotocografías.”

¿Con qué facilidad podría implementar estas explicaciones en su trabajo?

“Sencillo, muy sencillo. Creo nuevamente que la fortaleza está en él analizar desde lo normal y eso permitirá tener una mayor claridad. Partiendo de la base de lo normal voy a podemos analizar hasta la parte patológica con mayor claridad. Creo que es esto da bastante facilidad en la enseñanza en la conversación en él impartir en él compartir en realidad cómo son estos métodos con las diferentes personas del medio.”

Utilidad percibida

¿Qué tan útiles le parecieron las explicaciones obtenidas?

“Pues bastante útiles también, finalmente este es una parte en la que sin comparar ambos métodos terminan siendo bastante buenos en la toma de decisiones en realidad dejan muy claro que es lo normal y qué es lo patológico que es cuando debemos preocuparnos.”

¿Qué tan útil considera que sería implementar en su trabajo las explicaciones obtenidas?

“Pienso también que muy sencillo. Me gusta el que nos permita impartir quizás desde la base eso es lo que lo que más me gusta del método porque el partir de lo normal nos permite siempre saber qué es patológico y porqué tan grave estamos también en lo patológico. Entonces, me parece bastante útil el implementar este tipo de resultados. Evidentemente las interpretaciones para obtener resultados concéntricamente favorables a medida de lo posible, casi siempre. Ojalá fuera de esa forma, pero son recursos que nos ayudan de cierta manera a hacer un trabajo más óptimo más oportuno sobre todo, lo cual me parece muy importante.”

¿Qué recomendaciones daría para mejorar las explicaciones obtenidas?

“Bueno, nuevamente es bastante claro en realidad. Es igual, tendríamos muchas cosas más que estudiar la verdad. O sea es un sin fin de estos métodos son muchísimas cosas, pero para poder empezar y tener una base son bastante buenos los estudios que han hecho; son bastante claros y pienso que para empezar está muy muy bien. Se interpretan bien son fáciles de estudiar, me parecen fáciles de entender y por lo tanto eso va a permitir que tengan cierta accesibilidad al aplicar qué es lo más importante de los estudios. En realidad no tiene mucho sentido que podamos tener un material y no lo podamos aplicar. La finalidad de esto es que sea de una manera tan comprensible para todos y que nos permita aplicar en los diferentes grados.”

3.3 Conclusiones

- Respecto a las necesidades médicas planteadas en este trabajo de titulación, se concluye que es posible mediante metodologías de AI y XAI dar un soporte relevante en el campo de la medicina, ayudando a los especialistas a tomar decisiones que pueden ser difíciles considerando que cada una de estas recae en la responsabilidad de un criterio profesional.
- El algoritmo SHAP de XAI es una herramienta con gran capacidad explicativa que permite entender el comportamiento del algoritmo de Aprendizaje Automático para realizar una clasificación de una instancia, mostrando los valores que influyen para cada atributo, así como los atributos más influyentes.
- Se puede afianzar la confianza médica en el área de obstetricia para determinar la salud fetal con algoritmos de Aprendizaje Automático mediante el algoritmo SHAP, el cual realiza un recorrido de cada atributo en función de su valor e influencia para determinar por qué la salud del feto puede estar en riesgo y esto permite al médico especialista tomar una decisión temprana.
- Acorde a la revisión sistemática de literatura realizada en el presente trabajo de titulación, se concluye que es nula la literatura que estudie XAI en el área de obstetricia. Esto marca una importancia relevante de la investigación realizada ya que se explora un área de interés médico para apoyar la toma de decisiones en la salud fetal.
- Al realizar un análisis del conjunto de datos que se utilizó para el presente trabajo de titulación, se encuentran limitaciones importantes debido a que estos no tienen un balance de clases y acorde al objetivo planteado de explicabilidad no se puede realizar un sobre muestreo de datos, lo cual afecta directamente a la eficiencia del algoritmo de Aprendizaje Automático.
- Debido a la buena aceptación que tiene el estudio de XAI en el área de obstetricia por el especialista médico, se concluye que el presente trabajo de titulación tiene un fuerte potencial para trabajos futuros como una posible implementación que pueda ser utilizado por médicos con sus propios datos obtenidos de cardiotocografías y obtener una predicción de la salud fetal, así como explicaciones de esta.

3.4 Recomendaciones

- A pesar de que SHAP puede explicar cualquier modelo de AI que tenga una predicción de caja negra, el coste computacional de un Kernel Explainer resulta ineficiente al objetivo. Debido a esto es recomendable realizar explicaciones Locales agnósticas al modelo, caso contrario se pueden utilizar algoritmos de Aprendizaje Automático basados en arboles de decisión con los que podemos realizar Explicaciones basadas en los árboles para mejorar la eficiencia computacional.
- La revisión sistemática de literatura realizada en el presente trabajo de titulación fue de gran importancia para entender los estudios que se han realizado en XAI para la medicina, por lo cual es recomendable al interactuar con metodologías relativamente nuevas realizar un recorrido previo de la literatura existente, para tener una guía importante y poder determinar algoritmos eficientes de aprendizaje automático como los métodos más usados para realizar explicaciones confiables.
- A pesar de la buena aceptación médica obtenida en este trabajo de titulación, es recomendable contar con más especialistas médicos que interactúen con las explicaciones realizadas por el algoritmo de SHAP y encontrar posibles falencias del algoritmo cuando se trata de realizar explicaciones que puedan entender personas no informáticas. Con esto se podría mejorar la interpretabilidad del estudio y permitiría una posible futura implementación.

4 REFERENCIAS BIBLIOGRÁFICAS

- [1] M. van Lent, W. Fisher, and M. Mancuso, "An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior," in *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence*, 2004, pp. 900–907.
- [2] H. Hagras, "Toward Human-Understandable, Explainable AI," *Computer*, vol. 51, no. 9, pp. 28–36, 2018, doi: 10.1109/MC.2018.3620965.
- [3] I. Bratko, "Machine Learning: Between Accuracy and Interpretability," 1997.
- [4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.
- [5] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778 [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [6] J. Ochmann, S. Zilker, and S. Laumer, "Job Seekers' Artificial Intelligence-Related Black Box Concerns," in *Proceedings of the 2020 on Computers and People Research Conference*, 2020, pp. 101–102, doi: 10.1145/3378539.3393841 [Online]. Available: <https://doi.org/10.1145/3378539.3393841>
- [7] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
- [8] C. T. Wolf and K. E. Ringland, "Designing Accessible, Explainable AI (XAI) Experiences," *SIGACCESS Access. Comput.*, no. 125, Mar. 2020, doi: 10.1145/3386296.3386302. [Online]. Available: <https://doi.org/10.1145/3386296.3386302>
- [9] J. Tritscher, M. Ring, D. Schlr, L. Hettinger, and A. Hotho, "Evaluation of Post-hoc XAI Approaches Through Synthetic Tabular Data," 2020, pp. 422–430.

- [10] M. Reyes *et al.*, “On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities,” *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, May 2020, doi: 10.1148/ryai.2020190043.
- [11] M. Du, N. Liu, and X. Hu, “Techniques for Interpretable Machine Learning,” *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019, doi: 10.1145/3359786. [Online]. Available: <https://doi.org/10.1145/3359786>
- [12] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.
- [13] M. T. Keane and B. Smyth, “Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI),” 2020, pp. 163–178.
- [14] E. M. Kenny, E. D. Delaney, D. Greene, and M. T. Keane, “Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective,” 2021, pp. 20–34.
- [15] K. Kaczmarek-Majer, G. Casalino, G. Castellano, O. Hryniewicz, and M. Dominiak, “Explaining Smartphone-based Acoustic Data in Bipolar Disorder: Semi-Supervised Fuzzy Clustering and Relative Linguistic Summaries,” *Information Sciences*, 2021, doi: <https://doi.org/10.1016/j.ins.2021.12.049>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020025521012706>
- [16] E. Kalai and D. Samet, “On weighted Shapley values,” *International Journal of Game Theory*, vol. 16, no. 3, pp. 205–222, Sep. 1987, doi: 10.1007/BF01756292.
- [17] J. Dieber and S. Kirrane, “Why model why? Assessing the strengths and limitations of LIME,” Nov. 2020.
- [18] M. T. Keane and B. Smyth, “Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI),” 2020, pp. 163–178.
- [19] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-Wise Relevance Propagation: An Overview,” 2019, pp. 193–209.

- [20] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, Dec. 2018, pp. 81–87, doi: 10.1145/3278721.3278736.
- [21] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.*, vol. 1. 2021.
- [22] L. S. Shapley, "17. A Value for n-Person Games," in *Contributions to the Theory of Games (AM-28), Volume II*, Princeton University Press, 1953, pp. 307–318.
- [23] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," Feb. 2018.
- [24] S. L. Chau, J. Gonzalez, and D. Sejdinovic, "RKHS-SHAP: Shapley Values for Kernel Methods," Oct. 2021.
- [25] J. Spilka, G. Georgoulas, P. Karvelis, V. Chudáček, C. D. Stylios, and L. Lhotská, "Discriminating Normal from 'Abnormal' Pregnancy Cases Using an Automated FHR Evaluation Method," 2014, pp. 521–531.
- [26] I. Amer-Wåhlin *et al.*, "Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial," *The Lancet*, vol. 358, no. 9281, pp. 534–538, Aug. 2001, doi: 10.1016/S0140-6736(01)05703-8.
- [27] R. M. Grivell, Z. Alfirevic, G. M. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," *Cochrane Database of Systematic Reviews*, Sep. 2015, doi: 10.1002/14651858.CD007863.pub4.
- [28] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Jan. 2000.
- [29] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

- [30] F. J. García-Peñalvo, "Cómo hacer una Systematic Literature Review (SLR)." Zenodo, May 2021 [Online]. Available: <https://doi.org/10.5281/zenodo.4745223>
- [31] 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 Transactions on Medical Imaging*, vol. 40, no. 3, pp. 793–804, 2021, doi: 10.1109/TMI.2020.3036933.
- [32] M. Kiani, J. Andreu-Perez, H. Hagra, M. L. Filippetti, and S. Rigato, "A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Developmental Neuroscience," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020, pp. 1–8, doi: 10.1109/FUZZ48607.2020.9177711.
- [33] W. S. Liew, C. K. Loo, and 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.
- [34] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, and T. Wilaiprasitporn, "Parkinson's Disease Recognition Using SPECT Image and Interpretable AI: A Tutorial," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 22304–22316, Oct. 2021, doi: 10.1109/JSEN.2021.3077949.
- [35] A. Raison, P. Bourdon, C. Habas, and D. Helbert, "Explicability in resting-state fMRI for gender classification," in *2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME)*, Oct. 2021, pp. 5–8, doi: 10.1109/ICABME53305.2021.9604842.
- [36] D. O. Nahmias and 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," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2398–2406, doi: 10.1145/3394486.3403289 [Online]. Available: <https://doi.org/10.1145/3394486.3403289>
- [37] D. R. Chittajallu *et al.*, "XAI-CBIR: Explainable AI System for Content based Retrieval of Video Frames from Minimally Invasive Surgery Videos," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 66–69, doi: 10.1109/ISBI.2019.8759428.

- [38] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, and 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.
- [39] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghmi, and C. Fookes, "A Robust Interpretable Deep Learning Classifier for Heart Anomaly Detection Without Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2162–2171, 2021, doi: 10.1109/JBHI.2020.3027910.
- [40] S. Ghosh, P. Tino, and K. Bunte, "Visualisation and knowledge discovery from interpretable models," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206702.
- [41] P. A. Moreno-Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 4902–4910, doi: 10.1109/BigData50022.2020.9378460.
- [42] J. Duell, X. Fan, B. Burnett, G. Aarts, and S.-M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4, doi: 10.1109/BHI50953.2021.9508618.
- [43] G. Dong, Y. Ma, and 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.
- [44] A. Kumar, R. Manikandan, U. Kose, D. Gupta, and S. C. Satapathy, "Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 3s, Oct. 2021, doi: 10.1145/3457187. [Online]. Available: <https://doi.org/10.1145/3457187>
- [45] N. Potie, S. Giannoukakos, M. Hackenberg, and A. Fernandez, "On the Need of Interpretability for Biomedical Applications: Using Fuzzy Models for Lung Cancer Prediction with Liquid Biopsy," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6, doi: 10.1109/FUZZ-IEEE.2019.8858976.

- [46] 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.
- [47] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI: An Ontology-Based Approach to Black-Box Sequential Data Classification Explanations,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 629–639, doi: 10.1145/3351095.3372855 [Online]. Available: <https://doi.org/10.1145/3351095.3372855>
- [48] N. Keller, M. A. Jenny, C. A. Spies, and S. M. Herzog, “Augmenting Decision Competence in Healthcare Using AI-based Cognitive Models,” in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, Nov. 2020, pp. 1–4, doi: 10.1109/ICHI48887.2020.9374376.
- [49] A. Tahmassebi, J. Martin, A. Meyer-Baese, and A. H. Gandomi, “An Interpretable Deep Learning Framework for Health Monitoring Systems: A Case Study of Eye State Detection using EEG Signals,” in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 211–218, doi: 10.1109/SSCI47803.2020.9308230.
- [50] P. F. Khan and K. Meehan, “Diabetes prognosis using white-box machine learning framework for interpretability of results,” in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 2021, pp. 1501–1506, doi: 10.1109/CCWC51732.2021.9375927.
- [51] J. Kim, M. Kim, and Y. M. Ro, “Interpretation of Lesional Detection via Counterfactual Generation,” in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 96–100, doi: 10.1109/ICIP42928.2021.9506282.
- [52] N. Seedat, V. Aharonson, and Y. Hamzany, “Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning,” in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6, doi: 10.1109/CSDE50874.2020.9411529.
- [53] K. A. Thakoor, S. C. Koorathota, D. C. Hood, and P. Sajda, “Robust and Interpretable Convolutional Neural Networks to Detect Glaucoma in Optical Coherence

- Tomography Images,” *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2456–2466, 2021, doi: 10.1109/TBME.2020.3043215.
- [54] F. Stieler, F. Rabe, and B. Bauer, “Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1802–1809, doi: 10.1109/CVPRW53098.2021.00199.
- [55] C. K. Leung, D. L. x. Fung, D. Mai, Q. Wen, J. Tran, and J. Souza, “Explainable Data Analytics for Disease and Healthcare Informatics,” in *25th International Database Engineering & Applications Symposium*, 2021, pp. 65–74, doi: 10.1145/3472163.3472175 [Online]. Available: <https://doi.org/10.1145/3472163.3472175>
- [56] Q. Ye, J. Xia, and G. Yang, “Explainable AI for COVID-19 CT Classifiers: An Initial Comparison Study,” in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 521–526, doi: 10.1109/CBMS52027.2021.00103.
- [57] U. Pawar, C. T. Culbert, and R. O’Reilly, “Evaluating Hierarchical Medical Workflows using Feature Importance,” in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 265–270, doi: 10.1109/CBMS52027.2021.00075.
- [58] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-S², and L. Pereira-Leite, “Sisporto 2.0: A program for automated analysis of cardiocograms,” *The Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, Sep. 2000, doi: 10.1002/1520-6661(200009/10)9:5<311::AID-MFM12>3.0.CO;2-9.
- [59] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” *PLOS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
- [60] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [61] M. Wilke and V. J. Schmithorst, “A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data,”

- NeuroImage*, vol. 33, no. 2, pp. 522–530, Nov. 2006, doi: 10.1016/j.neuroimage.2006.07.010.
- [62] J. Miao and L. Niu, “A Survey on Feature Selection,” *Procedia Computer Science*, vol. 91, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111.
- [63] J. Li *et al.*, “Feature Selection: A Data Perspective,” *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, Nov. 2018, doi: 10.1145/3136625.
- [64] R. G. Shaw and T. Mitchell-Olds, “Anova for Unbalanced Data: An Overview,” *Ecology*, vol. 74, no. 6, pp. 1638–1645, Sep. 1993, doi: 10.2307/1939922.
- [65] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [66] D. Buschmann, C. Enslin, H. Elser, D. Lütticke, and R. H. Schmitt, “Data-driven decision support for process quality improvements,” *Procedia CIRP*, vol. 99, pp. 313–318, 2021, doi: 10.1016/j.procir.2021.03.047.
- [67] R. Noori, Z. Deng, A. Kiaghadi, and F. T. Kachoosangi, “How Reliable Are ANN, ANFIS, and SVM Techniques for Predicting Longitudinal Dispersion Coefficient in Natural Rivers?,” *Journal of Hydraulic Engineering*, vol. 142, no. 1, p. 04015039, Jan. 2016, doi: 10.1061/(ASCE)HY.1943-7900.0001062.
- [68] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for Explainable AI: Challenges and Prospects,” Dec. 2018.
- [69] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Quarterly*, vol. 13, no. 3, 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/1R70rkJCvocz1laO03Ubj9zAmeqicSsY?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/bryan_ortuno_epn_edu_ec/EefxRJ1kU2hFhDIqUPFsw9oBMnxJP6bKH9RTfdQqq3uD1g?e=DcKiPv

ANEXO III

Enlace dirigido a la matriz de la clasificación de los artículos para cada algoritmo de AI y XAI usados los diferentes estudios de la revisión sistemática de literatura:

https://epnecuador-my.sharepoint.com/:x/g/personal/bryan_ortuno_epn_edu_ec/ERz5DHB5EBhFmSEX9GBwUYcBc0MXH116JfJSg4-d9C0LxQ?e=M101Pg

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/bryan_ortuno_epn_edu_ec/Eeeleaz9fsJKmQ7jb0IJfe0BCKdzHfvswWU3dxgbaeB13A?e=0WC4xw