Escuela Politécnica Superior de Jaén
Máster Universitario en Ingeniería Informática

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior (Jaén)

Trabajo Fin de Máster

MÉTRICAS Y GUÍAS PARA EL DESARROLLO DE ALGORITMOS DE INTELIGENCIA ARTIFICIAL EXPLICABLE

Alumno/a: **Escobar Mimbrera, Ángela**

Tutores: Prof. D. Pedro González García
Prof. D. Mª José del Jesus Díaz
Dpto.: Informática

Septiembre, 2021

Universidad de Jaén
Escuela Politécnica Superior de Jaén
Departamento de Informática

Don Pedro González García y Doña María José del Jesus Díaz, tutor y cotutora del Trabajo Fin de Máster titulado: Métricas y guías para el desarrollo de algoritmos de Inteligencia Artificial Explicable, que presenta Ángela Escobar Mimbrera, autoriza su presentación para defensa y evaluación en la Escuela Politécnica Superior de Jaén.

Jaén, Septiembre de 2021

El alumno:

ESCOBAR MIMBRERA ANGELA - 26047453E
Firmado digitalmente por ESCOBAR MIMBRERA ANGELA - 26047453E
Fecha: 2021.09.06 09:22:12 +02'00'

Ángela Escobar Mimbrera

Los tutores:

GONZALEZ GARCIA PEDRO - 26210879X
Firmado digitalmente por GONZALEZ GARCIA PEDRO - 26210879X
Fecha: 2021.09.06 09:35:45 +02'00'

DEL JESUS DIAZ MARIA JOSE - 26485651R
Firmado digitalmente por DEL JESUS DIAZ MARIA JOSE - 26485651R
Fecha: 2021.09.06 09:59:00 +02'00'

Pedro González García
y María José del Jesus Díaz

Índice

Resumen

Una de las principales dificultades a la hora de desarrollar modelos de inteligencia artificial explicable, es que no existe una definición única, común y aceptada en la bibliografía específica sobre algunos conceptos y métricas que faciliten cuantificar la interpretabilidad de un modelo. En este trabajo se ha realizado una revisión de la terminología relacionada con inteligencia artificial explicable en la bibliografía especializada, que permite aportar definiciones integradoras, y proponer la unificación y agrupación de conceptos. Es un análisis inicial en un área que debe afrontar muchos retos, entre ellos el debate, consenso y propuesta sobre conceptos y métricas objetivas que cuantifiquen la explicabilidad de un modelo.

Palabras clave: inteligencia artificial explicable, XAI, explicabilidad, interpretabilidad, objetivos, propiedades, métricas

1. Introducción

En este capítulo se expone la motivación para la realización de un estudio bibliográfico sobre Inteligencia Artificial Explicable (XAI, del término en inglés e***X****plainable* ***A****rtificial* ***I****ntelligence*). Asimismo, se detallan los objetivos específicos fijados para alcanzar la definición de métricas y guías que permitan desarrollar métodos de XAI. Por último, se realizará una descripción de la organización de esta memoria de trabajo fin de máster.

1.1. Motivación

Cada vez son más los algoritmos de Inteligencia Artificial (IA) con los que convivimos día a día y que son capaces de decidir o facilitarnos la toma de decisiones habituales. Existen dos ejemplos muy claros, cuyo resultado repercute de una forma muy distinta en el usuario final. Por un lado, hay algoritmos de recomendación como los utilizados en plataformas tipo Netflix [Dye2020] para la recomendación de películas y/o series. Por otro lado, hay modelos de recomendación utilizados, por ejemplo, para ayudar en la toma de decisión sobre el diagnóstico de Covid19 [Ism2021]. La diferencia en el ámbito en el que se aplican los modelos de IA y la repercusión que estos tienen sobre el usuario según su recomendación final, es la que hace que en algunos casos (como el mencionado de la medicina, entre otros) sea necesario entender cómo determina el modelo dicha salida. De forma paralela, los modelos de IA desarrollados tienen un nivel de complejidad cada vez más alto, lo que dificulta proporcionar esta explicación. En este contexto surge la Inteligencia Artificial Explicable, XAI.

Antes de comenzar, y aunque se introducirá formalmente en el capítulo 2 de este trabajo, es necesario saber de qué se quiere decir cuando se habla de XAI. Pues bien, XAI es la forma de referirse a la explicabilidad de los métodos utilizados y los resultados obtenidos en IA. En la bibliografía especializada se han propuesto diferentes definiciones de XAI, por lo que que antes de seguir con este estudio, es positivo trabajar con una definición que permita aglutinarlas y unificarlas todas.

De igual forma a lo largo del tiempo se han introducido distintas descripciones (tanto dentro del ámbito del aprendizaje automático como fuera de él, como por ejemplo indica como por ejemplo indica W. James Murdoch et al. [Mur2019]), conceptos y terminología, relacionados con la explicabilidad, interpretabilidad, sencillez, etc. de los modelos. El análisis, debate y unificación de términos y conceptos puede contribuir al desarrollo de modelos en XAI.

Un modelo de XAI estará caracterizado por diferentes propiedades o características, que serán objetivos a alcanzar en el proceso de diseño de los mismos. Sobre ellas, de nuevo, existen diferentes visiones, propuestas y definiciones.

En este trabajo se realiza un análisis de conceptos en XAI que pretende ayudar a contribuir a la unificación de la terminología además de realizar un análisis de las métricas aplicables a uno de los tipos de modelos de IA más relacionados con la interpretabilidad, los modelos basados en reglas.

A continuación se presentan los objetivos marcados para cumplir con todo lo anterior.

1.2. Objetivos

El propósito general de este trabajo fin de máster es analizar y estudiar las propuestas existentes, determinar puntos coincidentes, solapados y/o contrapuestos entre ellas, y desarrollar una propuesta de unificación de las diferentes definiciones, conceptos y métricas utilizados por diferentes autores con referencia a XAI.

En base a estos propósitos se plantean los siguientes objetivos:

- Realizar una revisión bibliográfica sobre todo lo referente a XAI.
- Analizar las definiciones encontradas en la bibliografía especializada, y avanzar hacia la unificación de conceptos, propiedades y taxonomías existentes.
- Establecer una guía inicial que nos permita la definición de una taxonomía formal a partir del estudio de los diferentes conceptos, propiedades y objetivos de un modelo XAI.

- Revisar las métricas utilizadas para cuantificar la explicabilidad de un modelo centrándonos principalmente en los modelos de IA basados en reglas.

1.3. Estructura

Para alcanzar los objetivos planteados, esta memoria de investigación se estructura en los siguientes capítulos:

- **Capítulo 2.** En este capítulo se repasa la bibliografía en busca de las diferentes definiciones y acepciones del término de Inteligencia Artificial Explicable. Una vez revisadas estas, se trabajará sobre una definición unificada que englobe todo lo encontrado en la bibliografía, con el objetivo de proponer su uso de aquí en adelante.
- **Capítulo 3.** En este capítulo se realiza un análisis en la bibliografía especializada de toda la terminología XAI encontrada. Se estudian los diferentes conceptos que engloban estos modelos, iniciando el estudio con la agrupación de los conceptos, dependiendo de dónde se produzcan en el modelo, pasando después a la unificación de estos términos y terminando con una propuesta de definición para cada uno de ellos.
- **Capítulo 4.** Al igual que en el capítulo 3, durante este capítulo se revisa la terminología XAI encontrada fijando el foco en las propiedades y objetivos que deben estar presentes en un modelo XAI. Para ello se revisan todas las propiedades y objetivos que hacen que un algoritmo se pueda clasificar como explicable, unificando términos y realizando una propuesta de definición para cada uno de ellos.
- **Capítulo 5.** En este capítulo se realizará un repaso en la bibliografía especializada para analizar la forma de medir el nivel (o grado) de explicabilidad de los modelos de Inteligencia Artificial. Esto permite ver que se pueden encontrar dos tipos de medidas, medidas objetivas y medidas subjetivas. El capítulo termina analizando las métricas utilizadas para medir la explicabilidad de los modelos de IA basados en reglas.

- **Capítulo 6.** Finalmente, este capítulo señala las conclusiones y los resultados más destacados obtenidos durante el proceso de investigación y análisis, así como las líneas de trabajo futuras.

La memoria concluye con la **Recopilación Bibliográfica** delas fuentes más destacadas de la materia estudiada.

2. Definición de Inteligencia Artificial Explicable

Antes de avanzar, se debe comenzar por saber qué significa hablar de XAI. ¿Qué es XAI? XAI es el acrónimo de Explainable Artificial Intelligence o lo que es lo mismo, XAI es la forma de referirse a la explicabilidad de los métodos utilizados en inteligencia artificial. XAI no sólo se centra en la explicación de los métodos, algoritmos o técnicas utilizadas, sino que también hace referencia a la explicación de los resultados hacia el usuario final.

En la bibliografía especializada se han propuesto diferentes definiciones de XAI. Antes de entrar a realizar una revisión de los conceptos y propiedades que deberían de incluir los algoritmos XAI, es importante analizar las propuestas de definición del área existentes y determinar una propuesta de definición que sea capaz de englobar y hacer lo más extensible posible el significado de XAI. En este capítulo, se hace un repaso de las distintas definiciones y acepciones encontradas en la bibliografía acerca de este término. Se concluye aportando una definición que engloba todo lo encontrado y sirva para sentar las bases del estudio.

2.1. Diferencias bibliográficas en una misma definición

Una vez revisada la bibliografía sobre el tema, se descubre que muchos autores hacen una breve mención a realizar algoritmos de inteligencia artificial explicable, es más, indican que es importante que la IA evolucione en algunos sectores a algoritmos XAI que permitan explicar la toma de decisiones en los algoritmos. Sin embargo, la mayoría de ellos no dan una definición clara de XAI, o si lo hacen, es una definición parcial, que no permite englobar todo lo necesario para llegar a una definición completa del término. En la siguiente gráfica (ilustración 1) podemos ver la evolución del número de apariciones del término XAI en revistas indexadas:

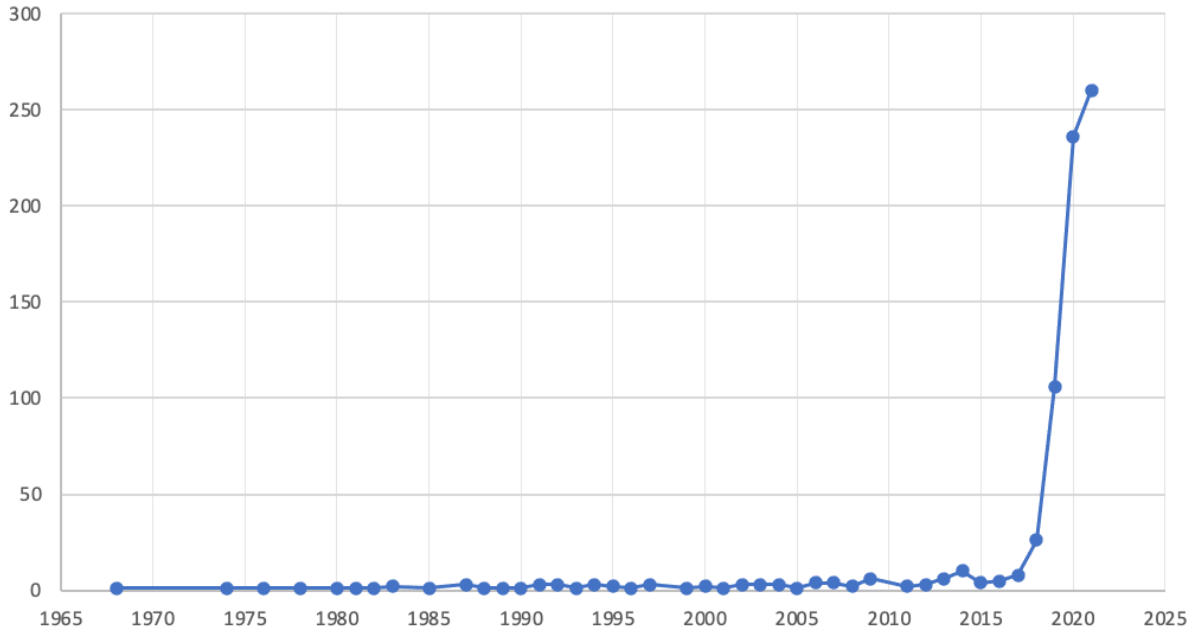

Ilustración 1. Evolución de la aparición del término XAI en la bibliografía especializada. Fuente de datos: Scopus

La primera referencia bibliográfica en la que se encuentra el término XAI y se define de forma completa es del año 2004 y viene dada por Van Lent et al. [Len2004]. En ella se indica que XAI es un campo de investigación que tiene como objetivo hacer más entendibles a los humanos los resultados obtenidos por sistemas IA. El término se acuñó como tal para poder describir la nueva técnica utilizada en su sistema.

Más adelante, en 2016, Gunning [Gun2016] comienza a hablar de explicabilidad, pero en su artículo no se indica una definición como tal sino que señala que esta será esencial para que los usuarios comprendan y confíen en la nueva generación de modelos emergentes de Inteligencia Artificial. Gunning et al. [Gun2018], un poco más adelante, comienzan a dar una visión más cercana a lo que serían algoritmos XAI pero tampoco llegan a definirla. En este artículo indican que es conveniente brindar al usuario final explicaciones que le permitan comprender las fortalezas y debilidades del sistema generado, así como la toma de decisión que más adelante le permita realizar una corrección de errores en el caso de que sea necesario.

De esta forma, Gunning et al. [Gun2018] y Van Lent et al [Len2004], avanzan en su concepto de XAI pero aún no dan una definición, como tal, del mismo.

Otros ejemplos en los que también se habla del término explicabilidad, pero sin aportar una definición concreta del mismo, se encuentran en el artículo de Barocas et al. [Bar2018] donde hacen una apreciación a que los algoritmos de decisión deben garantizar que, tanto la toma de decisión como los datos que impulsen esa toma, deben poder ser explicados al usuario final sin necesidad de utilizar lenguaje técnico.

Por otra parte, la comunidad FICO (*Fair Isaac Corporation*) [FIC2018] tras organizar el *XML Challenge*, ve XAI como una oportunidad o desafío que permite abrir los diferentes modelos de caja negra y que, a partir de técnicas, estos modelos sean capaces de precisar y proporcionar una explicación confiable que satisfaga las necesidades de los usuarios.

Otro ejemplo, en este caso entendiéndolo como definición dentro del contexto de los sistemas de aprendizaje automático (*Machine Learning*), es el dado por Doshi-Velez et al. [Dos2017] que lo define como la capacidad de explicar o presentar en términos comprensibles para un ser humano.

Como se ha visto hasta ahora, aunque varios autores hablan de la explicabilidad de los modelos y su necesidad para confiar en los mismos, no se ha encontrado una definición del término que sea capaz de englobar tanto la parte de explicabilidad procedente al algoritmo como la referente a la presentación de los resultados. Lo mismo ocurre con el término dado por Barredo et al. [Bar2020] donde se define XAI de la siguiente manera:

> *"Dada una audiencia, una Inteligencia Artificial Explicable es aquella que produce detalles o razones para que su funcionamiento sea claro o fácil de entender".*
>
> Barredo et al. [Bar2020]

Una vez revisada toda esta bibliografía, podemos dar una nueva definición de XAI. Esta engloba tanto la explicación de los resultados para hacerlos comprensibles al tomador de decisiones como la explicación de los métodos y algoritmos utilizados.

2.2. Unificación de terminología para una única definición

Como se ha visto en la sección anterior, son muy pocos autores los que dan una definición clara de que es XAI, aunque muchos de ellos sí que la mencionan y la ven necesaria.

Todos ellos mencionan la Inteligencia Artificial Explicable como aquella que es capaz de dar al usuario final una explicación en la toma de decisión realizada por el algoritmo de inteligencia artificial, pero, ¿cómo proporcionamos esta explicación al usuario final? Sería conveniente tener en cuenta el conocimiento sobre la materia que tenga el usuario final, así como la elección del formato en el que se va a dar esa explicación. Mosheni et a. en [Mos2020] presentan una segregación de usuarios con respecto al nivel de conocimiento al igual que se distinguen distintos formatos (formato visual mediante ilustraciones, gráficos… o formato texto a partir del lenguaje natural…) a la hora de representar la toma de decisiones.

Partiendo de la definición dada por Barredo et al. [Bar2020], y la segregación de usuarios y formatos que se describe en Mosheni et al. [Mos2020], se ha llegado a la siguiente definición:

> “Se entiende que un algoritmo de inteligencia artificial explicable es capaz de producir detalles o razones en cada una de sus partes que haga entender y esclarecer su funcionamiento, así como que consiga dar una explicación de porqué se deriva a una toma de decisión final, dependiendo del tipo de usuario y formato al que llega esa explicación.”

A continuación, se explica esta definición de forma visual. Para ello se utiliza como apoyo la definición dada por Gunning et al. [Gun2016].

En la ilustración 2 se puede ver como un sistema XAI está formado por el modelo y por la interfaz explicable con la que el tomador de decisiones interactúa para tomar una decisión a partir de los resultados de este. Se empieza a ver como se da cabida a la forma en la que el modelo da sus resultados a partir de la interfaz explicable.

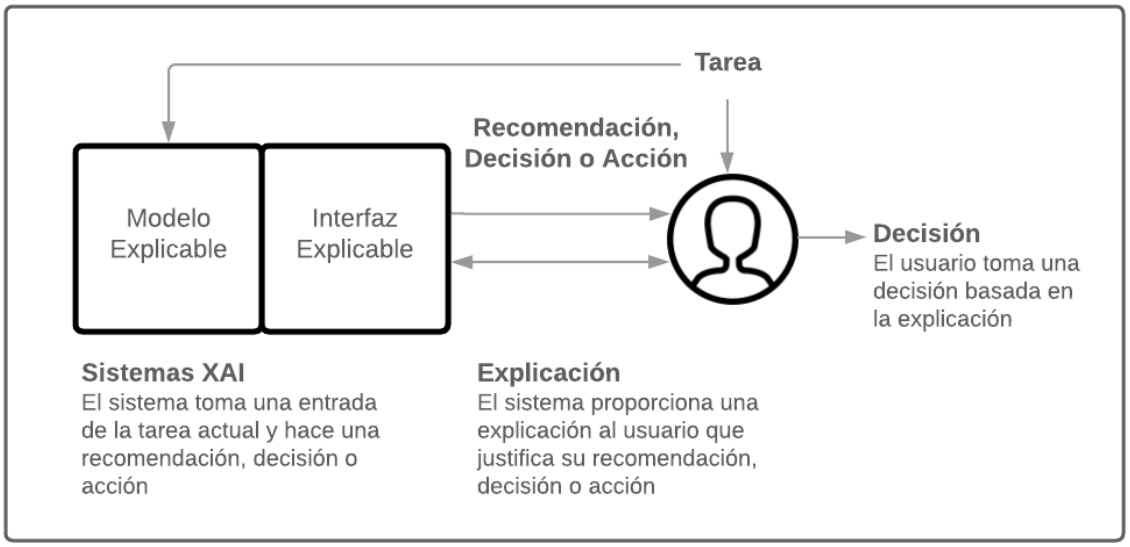

Ilustración 2. Definición visual de XAI
Fuente: [Gun2016]

3. Conceptos en Inteligencia Artifical Explicable

En el capítulo anterior se ha dado una definición de XAI. En este capítulo se revisan las diferentes ideas o conceptos que ayudan a construir y entender aún más la definición antes propuesta.

En la bibliografía especializada existen discrepancias a la hora de definir los conceptos relacionados con XAI, por lo que en el desarrollo de este capítulo se hace un repaso de los diferentes conceptos que proponen los distintos autores.

Tras un repaso exhaustivo de la bibliografía se observa que los términos asociados a XAI que aparecen con mayor frecuencia son los siguientes:

- Inteligibilidad *(understandability),*
- Comprensibilidad *(comprehensibility),*
- Transparencia *(transparency),*
- Explicabilidad *(explainability),*
- Interpretabilidad *(interpretability) y*
- Explicabilidad Post-Hoc *(post-hoc explainability).*

Estos conceptos se pueden categorizar en función de su aplicabilidad al modelo o al resultado obtenido por el mismo, y se analizarán en base a esta clasificación. Para ello se estudian las diferentes definiciones encontradas en la bibliografía, comenzando siempre por la aportada por la Real Academia Española de la Lengua, en adelante RAE. Tras el estudio de términos y conceptos, se concluye con una propuesta de definición propia del concepto que unifica términos diferentes utilizados con frecuencia en la bibliografía especializada para hacer referencia al mismo concepto. El objetivo, como se mencionaba al comienzo, es unificar conceptos en el área de investigación de XAI.

3.1. Conceptos aplicables al modelo

En esta sección se estudian los conceptos relativos a la explicación de un modelo a partir de su algoritmo y funcionalidad: Inteligibilidad, comprensibilidad y transparencia.

3.1.1. Inteligibilidad

Revisando el concepto inteligibilidad en la RAE, esta nos indica que es “Cualidad de inteligible”, es decir, inteligibilidad es aquello que puede ser entendido. Aparte de la definición académica del adjetivo, distintos autores han introducido diferentes definiciones en el ámbito de la Inteligencia Artificial.

Según Barredo et al. [Bar2020], un modelo XAI es inteligible si el mismo contiene la característica que haga entendible o comprensible para un ser humano cómo funciona el modelo sin necesidad de explicación del funcionamiento interno del algoritmo utilizado.

En la misma línea que la definición anterior se encuentra el concepto de inteligibilidad en el artículo de Montavon et al. [Mon2018], que indica que un modelo es inteligible cuando es capaz de dar una comprensión funcional del mismo sin llegar a profundizar en las diferentes capas del modelo para que este sea entendible.

También esta el caso de Lipton [Lip2017] en cuyo estudio señala que hay otros autores en diferentes artículos que igualan el concepto de inteligibilidad con el de interpretabilidad. Más adelante se estudiará la definición de interpretabilidad pero, como introducción, lo que estos autores indican con esta igualación de conceptos es que para ellos, que un modelo sea inteligible es que de por sí ya es interpretable.

Una vez revisadas las diferentes definiciones dadas por los autores en la bibliografía consultada para Inteligibilidad, se puede establecer que:

> “Un modelo es inteligible si funcionalmente es entendible por un ser humano sin necesidad de profundizar en el funcionamiento interno mismo.”

3.1.2. Comprensibilidad

Otro de los conceptos utilizados para poder definir XAI es comprensibilidad. Para este concepto se han encontrado diferentes autores que lo referencian.

La RAE indica que comprensibilidad es “Cualidad de comprensible”, por tanto, que se puede comprender. Ahondando un poco más, una de las acepciones que incluye la RAE para comprender es: “Entender, alcanzar o penetrar algo”. Vista esta definición, se observa que la definición entre este concepto e Inteligibilidad, visto anteriormente, es muy similar. Se revisan ahora las diferentes referencias encontradas en la bibliografía específica para poder comprender si existen matices diferenciadores para ellos.

Una de las primeras referencias sobre comprensibilidad la aporta Michalski [Mic1983] que dice que “los resultados de la inducción por computadora deben ser descripciones simbólicas de entidades dadas, semántica y estructuralmente similares a las que un experto humano podría producir observando las mismas entidades. Donde los componentes de estas descripciones deben ser comprensibles de forma individual y directamente interpretables en lenguaje natural”. En esta definición Michalski ya anticipa la importancia de que la resolución sea comprensible para el usuario que va a recibir esa información.

Partiendo de esta premisa también se encuentra lo expuesto por Freitas [Fre2014] que indica que la importancia de este concepto es subjetiva ya que depende de los usuarios y del dominio de la aplicación. Además, en este artículo, Freitas realiza un exhaustivo análisis de “cuánto de comprensibles” son los diferentes modelos de clasificación. Este análisis lo realiza para poder desmentir la afirmación de que mientras más profundidad tenga un modelo menor será su comprensibilidad, llegando a indicar que se debe trabajar para que la precisión de un modelo no se vea disminuida a medida que se incrementa su comprensibilidad.

Otros autores que hacen referencia a la comprensibilidad son Guidotti et al. [Gui2018] que igualan el concepto de interpretabilidad y el de comprensibilidad. Es la primera ocasión en que dos investigadores utilizan la misma definición para estos dos términos.

Por último, otros autores que hacen referencia a este concepto son Fernandez et al. que indican que “la comprensibilidad se refiere a la capacidad de un algoritmo de aprendizaje para representar su conocimiento aprendido de manera comprensible para el ser humano” [Fer2019]. Esta misma definición es recogida posteriormente por Barredo et al. [Bar2020].

Revisadas todas las referencias encontradas en la bibliografía específica, se ha llegado a la conclusión de que

> “Un modelo es comprensible si tiene suficiente capacidad para representar de forma entendible para el ser humano, que lo esté analizando, el conocimiento aprendido sin perjudicar en su eficiencia sea cual sea la profundidad del mismo.”

Dada esta definición, se observa que el matiz que diferencia Inteligibilidad de Comprensiblidad y que hace que las dos sean necesarias para su definición, se halla en hasta qué profundidad algorítmica del modelo este es capaz de ser entendible por un ser humano sin que se llegue a perjudicar la eficiencia del mismo.

3.1.3. Transparencia

El siguiente concepto XAI a revisar es el de Transparencia.

La RAE indica que Transparencia es algo que tiene la cualidad de transparente. Al seguir profundizando, se observa que dentro de las definiciones dadas para transparente, aparece: “Claro, evidente, que se comprende sin duda ni ambigüedad.” Utilizando esta definición como guía, se determina que para que un modelo sea transparente, a su vez debe ser comprensible. Se analiza a continuación qué indican los diferentes autores sobre este concepto.

La primera definición encontrada es la facilitada por Z.C. Lipton [Lip2017], e indica lo siguiente:

> "Informalmente, la transparencia es lo opuesto a la opacidad. Connota un cierto sentido de comprensión del mecanismo por el cual el modelo funciona".
>
> Z.C. Lipton [Lip2017]

Dada esta definición, a continuación, el mismo autor [Lip2017] indica los distintos niveles de transparencia que se pueden dar en un modelo: "Consideramos la transparencia a nivel del modelo completo (simulabilidad, en inglés simulatability), a nivel de los componentes individuales (descomponibilidad, en inglés decomposability) y a nivel del algoritmo de entrenamiento (transparencia algorítmica, en inglés *algorithmic transparency*)". Es la primera vez que dentro de un concepto encontramos niveles en los que se pueda aplicar. En este caso, la transparencia se puede dar en tres niveles:

- el modelo puede ser totalmente transparente en todos sus procesos,
- sólo ser transparente de forma modular a partir de la descomposición del mismo, o
- que sólo lo sea su algoritmo de aprendizaje.

La siguiente definición de transparencia la establece Barredo et al. [Bar2020], e indica que:

> "Un modelo se considera transparente si por sí mismo es comprensible".
>
> Barredo et al. [Bar2020]

Una vez dada esta definición, estos autores hacen referencia a la misma descomposición por niveles de la transparencia de un modelo planteada por Z.C. Lipton [Lip2017].

Visto todo lo anterior, se llega a la conclusión de que:

> "Un modelo es transparente cuando su modelo funcional es comprensible por sí mismo. Si se utilizan los niveles de transparencia indicados Z.C. Lipton [Lip2017], la cuestión a plantear ahora sería la siguiente: ¿se considera que un modelo es totalmente transparente si cumple el nivel de simulabilidad o se considera un modelo transparente sólo con cumplir uno de los tres niveles? Para responder a esta pregunta, hay que centrarse en si se está tratando con un modelo de caja blanca o caja negra. Si se trata de un modelo de caja negra -como son por ejemplo las redes neuronales-, se considerará transparente sólo si su algoritmo de entrenamiento es comprensible funcionalmente (transparencia algorítmica). En el caso de un modelo de caja blanca, para considerar el modelo transparente, debe cumplir la transparencia a nivel de modelo completo (simulabilidad), es decir, debe ser transparente a nivel de descomponibilidad y de transparencia algorítmica."

Una vez revisados estos conceptos y dada una definición de cada uno de ellos, la siguiente cuestión a plantear es: ¿Se podría establecer una gradación de mayor a menor explicabilidad en los conceptos aplicables a nivel de modelo? Partiendo de las definiciones dadas para cada uno de estos modelos, se observa que su matiz diferenciador es la profundidad modular o algorítmica en la que cada uno de estos es capaz de ser entendible por el ser humano.

Por tanto, se llega a la siguiente gradación de estos conceptos de menor a mayor de la siguiente forma (Ilustración 3):

- Inteligibilidad
- Comprensibilidad
- Transparencia
    - Simulabilidad
    - Descomponibilidad
    - Transparencia Algorítmica

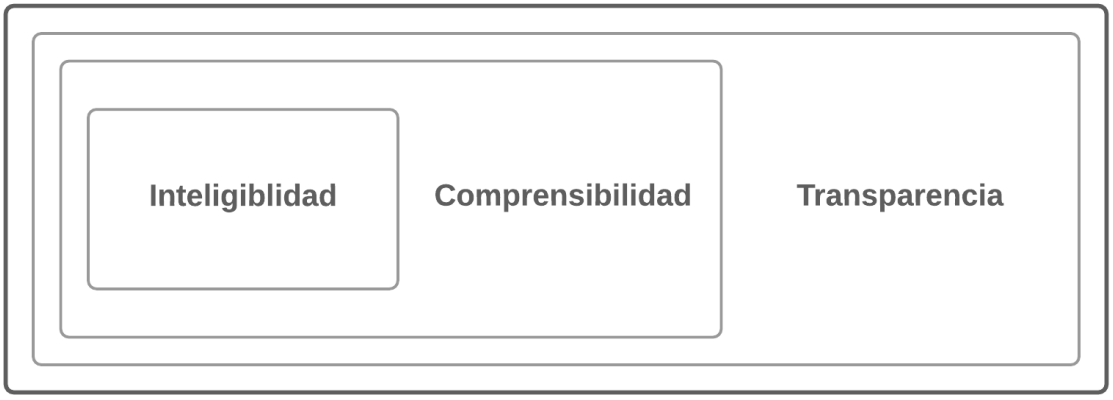

Ilustración 3. Gradación de los conceptos aplicables al modelo

En base a lo analizado para términos XAI relativos a modelos, la transparencia, con sus diferentes niveles (simulabilidad, descomponibilidad y transparencia algorítmica -el cual representa el valor más alto de transparencia), representa un mayor nivel de entendimiento en XAI, seguida de la comprensibilidad, y por último de la inteligibilidad. En todo caso, en función del objetivo final del modelo de IA, será más útil considerar el modelo XAI desde la perspectiva ofrecida por uno u otro término. Por indicar un ejemplo, no siempre será necesario tratar el modelo desde la perspectiva de la transparencia algorítmica, en ocasiones será suficiente tratar el modelo para que el usuario confíe en que es comprensible.

3.2. Conceptos aplicables a los resultados del modelo

En esta sección se analizan los conceptos que sirven para explicar los resultados obtenidos por un modelo, es decir, la salida de los mismos y cómo se representa. Los conceptos incluidos en esta categoría son: explicabilidad o interpretabilidad e interpretabilidad post-hoc.

3.2.1. Explicabilidad

Se comienza por revisar las diferentes definiciones vistas para Explicabilidad.

Primero, se busca la definición del concepto en la RAE. Como no se obtienen resultados para explicabilidad, se analiza el significado del adjetivo “explicable” para la RAE: “Que se puede explicar”. Profundizando en la búsqueda, se analiza el término “explicar”, entre cuyas definiciones se encuentra la siguiente, que se adapta a esta materia: “Declarar o exponer cualquier materia, doctrina o texto difícil, con palabras muy claras para hacerlos más perceptibles”.

Se revisan ahora las definiciones dadas por los diferentes autores en la bibliografía específica sobre este concepto. Se parte de la definición que indican Adadi et al. [Ada2018] en su artículo, en el que indican que la explicabilidad está estrechamente relacionada con el concepto de interpretabilidad, ya que “los sistemas interpretables son explicables si sus operaciones pueden ser comprendidas por el ser humano.” Además, en el mismo artículo, realizan una breve observación sobre el número de veces que se hace referencia a las palabras explicabilidad e interpretabilidad dentro de la comunidad de aprendizaje automático y nos vienen a decir que “interpretable” se llega a utilizar muchas más veces en esta comunidad que “explicable” refiriéndose al mismo significado.

Vista esta definición en la cual se igualan los términos interpretabilidad y explicabilidad, se pasa a revisar la definición dada por los autores Montavo et al. [Mon2018] sobre cómo ha de ser la explicabilidad, en la que indican que:

> "una explicación es el conjunto de características del dominio interpretable, que han contribuido para un ejemplo dado a producir una decisión."
>
> Montavo et al. [Mon2018]

Estos autores vuelven a hablar del dominio interpretable que utilizaban para poder definir el concepto de interpretabilidad y asocian este dominio a la explicación que permite esa toma de decisiones. Ese dominio, puede venir dado a partir de una visualización gráfica, por ejemplo, un punto culminante en un mapa de calor, o dentro del procesamiento del lenguaje natural, como forma de texto resaltado.

La última definición que se estudia sobre explicabilidad en este trabajo, es la dada por Barredo et al. [Bar2020]. Estos autores indican que:

> "la explicabilidad está asociada con la noción de explicación como una interfaz entre los seres humanos y un responsable de la toma de decisiones que es, al mismo tiempo, una representación exacta del responsable de la toma de decisiones y comprensible para los seres humanos".
>
> Barredo et al. [Bar2020]

Lo que viene a decir, es que para que el resultado de un modelo pueda cumplir el concepto de explicabilidad, debe tener la característica de poder ser comprensible tanto para el tomador de decisiones que tiene que realizar la toma de una decisión a partir del resultado como para el humano que debe de entender el porqué de esa toma de decisión.

3.2.2. Interpretabilidad

Se repasan ahora las diferentes definiciones dadas para el concepto de "Interpretabilidad".

Según la RAE interpretar es "Explicar acciones, dichos o sucesos que pueden ser entendidos de diferentes modos."

Revisando diferentes artículos, se observa que todos ellos llegan a la misma conclusión. Por ejemplo, Doshi-Velez et al. [Dos2017] definen la interpretabilidad como "la habilidad de explicar o presentar en términos comprensibles a un humano".

Esta misma definición la recogen en sus diferentes artículos los autores Guidotti et al. [Gui2018] y Barredo et al. [Bar2020].

Además, como se ha comentado en el apartado anterior, los autores Guidotti et al. [Gui2018], **equiparan** los términos de comprensibilidad e interpretabilidad ya que para ellos, ambos no se entienden por separado sino que uno engloba al otro, es decir, que un modelo si es interpretable es de por sí comprensible.

Otra definición es la dada por Montavon et al. [Mon2018]. En ella indican que "una interpretación es el mapeo de un concepto abstracto en un dominio que el humano puede entender". A partir de esta definición, se da a entender que el resultado se puede dar en diferentes contextos y formatos que se adapten al conocimiento de la persona que los revise.

3.2.3. Unificación de conceptos. Explicabilidad

Una vez revisadas tanto la definición de Interpretabilidad como la de Explicabilidad, se propone la unificación de estos términos bajo el concepto de **Explicabilidad**. Esa propuesta de unificación de conceptos se base en los artículos de Guidotti et al. [Gui2018] y de Adadi et al. [Ada2018]. En el primero se realiza una igualación de términos, mientras que en el segundo se habla de una relación estrecha entre ambos.

Es por ello que este estudio propone utilizar el concepto de **Explicabilidad**, siendo esta la

> "Habilidad de un modelo de explicar en términos comprensibles tanto para un ser humano como para un responsable de la toma de decisiones, la salida del modelo."

Dicha explicación se puede realizar en diferentes formatos pero siempre teniendo en cuenta el dominio en el que el tomador de decisiones se encuentre.

3.2.3.1. Explicabilidad Post-Hoc

El último concepto que se revisa dentro de los conceptos relacionados con los resultados del modelo es el de Explicabilidad Post-Hoc. Este concepto se considera como un subtipo dentro del concepto de Explicabilidad.

Este concepto o enfoque de la explicabilidad de un modelo no se centra en la explicación del funcionamiento del modelo sino en su salida, confiriendo información útil para el humano al cual llega ese resultado. Una de las ventajas fundamentales de este tipo de explicabilidad es "que se pueden interpretar modelos opacos a posteriori, sin sacrificar el rendimiento predictivo del modelo" [Lip2017].

Otro de los autores que referencian este concepto es Barredo et al. [Bar2020], en cuyo artículo se indica que esta "se dirige a los modelos que no son fácilmente interpretables por diseño".

En ambos artículos se establecen las diferentes formas en las que se pueda dar esta interpretabilidad:

- a través del lenguaje natural,
- a través de la explicación a partir de visualizaciones gráficas,
- a través de la explicación por aproximación local y

- a través de la explicación por relevancia de características o de ejemplo. Un ejemplo de la explicación por relevancia de características la podemos encontrar en una radiografía médica de un tumor, este se puede clasificar como maligno porque para el modelo se parece o aproxima mucho a otros tumores de este tipo. [Lip2017]

Una vez vistas ambas definiciones y las formas en las que se puede dar, se propone que:

> "la Explicabilidad Post-Hoc es aquella que permite interpretar y/o explicar modelos opacos o poco interpretables una vez que este ha devuelto sus resultados sin que se penalice el rendimiento predictivo del mismo."

Esta explicabilidad Post-Hoc se puede dar a partir de las formas ya descritas anteriormente.

3.3. Resumen de conceptos XAI

Una vez que se ha hecho la revisión de la bibliografía, se han agrupado los conceptos y se ha dado una definición para cada uno de ellos, se presenta una tabla resumen con todo lo descrito anteriormente:

Tipo	Concepto	Definición	Referencias
Aplicables al modelo	Inteligibilidad	Un modelo es inteligible si funcionalmente es entendible por un ser humano sin necesidad de profundizar en el funcionamiento interno del mismo.	RAE [Bar2020] [Mon2018] [Lip2017]
	Comprensibilidad	Un modelo es comprensible si tiene suficiente capacidad para representar de forma entendible para el ser humano, que lo esté analizando, el conocimiento aprendido sin perjudicar en su eficiencia sea cual sea la profundidad del mismo.	RAE [Mic1983] [Fre2014] [Gui2018] [Fer2019] [Bar2020]
	Transparencia	Un modelo se considera transparente si por sí mismo es comprensible. Se definen distintos niveles de transparencia dependiendo del tipo de modelo de la siguiente forma. Si se trata de un modelo de caja negra, lo consideraremos transparente si somos capaces de llegar a una transparencia algorítmica del mismo en el algoritmo de entrenamiento. En el caso de un modelo de caja blanca, para considerar el modelo transparente, debe de cumplir la transparencia a nivel de modelo completo, es decir, debe de ser transparente a nivel de descomponibilidad y de transparencia algorítmica.	RAE [Lip2017] [Bar2020]
Aplicables a los resultados	Explicabilidad	Habilidad de un modelo de explicar tanto en términos comprensibles tanto para un ser humano como para un responsable de la toma de decisiones, la salida del modelo.	RAE [Ada2018] [Mon2018] [Bar2020] [Dos2017] [Gui2018]
	Explicabilidad Post-Hoc	Es aquella que nos permite poder interpretar y/o explicar modelos opacos o poco interpretables una vez este ha devuelto sus resultados sin que se penalice el rendimiento predictivo del mismo.	[Lip2017] [Bar2020]

Tabla 1. Resumen de los conceptos aplicables a XAI

4. ¿Qué se espera de un modelo de Inteligencia Artificial Explicable? Propiedades y objetivos de un modelo XAI

Una vez revisados los conceptos relacionados con XAI, se analizan los diferentes términos relacionados con las propiedades u objetivos que ha de tener un modelo XAI, comenzando con la revisión de la bibliografía especializada y concluyendo con un resumen de todo lo revisado.

4.1. Propiedades y objetivos de un modelo XAI

El análisis de estos términos en la bibliografía especializada permite comprobar que hay autores que etiquetan cada uno de ellos de forma distinta aunque en su definición llegan a puntos comunes. Estos términos corresponden tanto a características deseables de los modelos de XAI como a objetivos que se pretenden alcanzar con estos modelos, que deberán incorporarse en sus mecanismos de evaluación.

Las propiedades y objetivos que se revisarán son los siguientes:

- Privacidad *(privacy)*
- Fiabilidad *(reliability)*
- Causalidad *(causality)*
- Transferibilidad *(transferability)*
- Informatividad *(informativeness)*
- Imparcialidad *(fairness)* e
- Interactividad *(interactivity)*

Para ello se seguirá el mismo esquema utilizado en el capítulo anterior. Se comienza por la definición encontrada en la Real Academia Española, continuando con las diferentes acepciones encontradas en la bibliografía revisada, y concluyendo con una propuesta de definición propia de la propiedad u objetivo estudiado.

4.1.1. Privacidad

El primer término a revisar será el de Privacidad. Para la RAE la privacidad es el ámbito de la vida privada que se tiene derecho a proteger de cualquier intromisión. Al seguir buscando, en este caso, privado, una de sus definiciones, es particular y personal de cada individuo. Aunque en primera instancia no se observa que lo encontrado tenga unión con las propiedades XAI, se revisan las diferentes definiciones de esta propiedad para ver cómo se puede relacionar.

La primera referencia encontrada es el artículo de Doshi-Velez et al. [Dos2017], en el que se indica que la privacidad significa que el método protege la información sensible de los datos.

Otros de los autores que hablan de este término son Barredo et al. [Bar2020]. En este artículo hablan de la sensibilización o concienciación de la privacidad (privacy awareness) e indican que tanto el no ser capaces de entender lo que ha sido capturado por el modelo como el almacenado en su representación interna puede suponer una violación de la privacidad. Por el contrario, también ocurre que la posibilidad de explicar las relaciones internas de un modelo entrenado por terceros no autorizados también puede comprometer la privacidad diferencial del origen de datos.

En conclusión, volviendo a las primeras definiciones revisadas de la RAE, estas tienen sentido si se está hablando de los datos que trata el modelo, datos sensibles que tienen que ver con la persona individual. Por tanto:

> “para que un modelo cumpla la propiedad de privacidad, este debe de ser capaz de proteger toda la información sensible que se maneje en cada una de sus etapas, independientemente de que el modelo sea entrenado por nosotros mismos o por terceros, siendo particularmente riguroso en las etapas de explicación del modelo para evitar exponer la privacidad de los datos sensibles al intentar hacer más explicable el modelo”.

4.1.2. Fiabilidad

El siguiente término a estudiar será el de fiabilidad. Diferentes autores hablan de esta propiedad refiriéndose a ella con diferentes acepciones, que una vez definidas tienen a mostrar un significado muy similar. Es por esto que se van a aunar todas estas definiciones en una misma propiedad. Al igual que en el resto de los términos, se comienza por la RAE, la cual nos indica que se entiende la fiabilidad como un método o modelo que ofrece seguridad o buenos resultados.

La primera definición encontrada en la bibliografía especializada está en el artículo de Doshi-Velez et al. [Dos2017]. En él se habla de la fiabilidad y la robustez, *reliability and robustness*, como propiedades que determinan si los algoritmos pueden alcanzar ciertos niveles de rendimiento ante la variación de las entradas o de los parámetros.

La siguiente definición se encuentra en Lipton [Lip2017]. En ella se habla de la propiedad de la confianza, *trust,* e indica que si se entiende la confianza como la fiabilidad de que un modelo debe funcionar correctamente, entonces, este sólo será confiable cuando sea suficientemente preciso. Por otra parte, este mismo autor, considera que esta propiedad es subjetiva, ya que también podríamos entender que un modelo es confiable o fiable si cuando los objetivos de formación y despliegue difieren. Por tanto, la confianza podría denotar la seguridad de que el modelo funcionará con respecto a los objetivos y escenarios reales.

Las siguientes definiciones se encuentran en el artículo de Barredo et al. [Bar2020] en el que se diferencia entre fiabilidad, *trustworthiness,* y confianza, *confidence*. Los autores consideran la fiabilidad como la confianza de que un modelo actuará como se pretende al enfrentarse a un problema determinado. Aunque ciertamente debería ser una propiedad de cualquier modelo explicable, no implica que todo modelo digno de confianza pueda considerarse explicable por sí mismo, ni que la fiabilidad sea una propiedad fácil de cuantificar. La confianza puede estar lejos de ser el único objetivo de un modelo explicable, ya que la relación entre ambos, si se acuerda, no es recíproca.

Vista esta definición, se analiza ahora la definición dada en este artículo para confianza, que se define como la generalización de la robustez y la estabilidad, la confianza debe evaluarse siempre sobre un modelo en el que se espera fiabilidad.

Revisadas las definiciones encontradas en la bibliografía especializada, así como la definición dada por la RAE, se propone una definición unificada de este término de la siguiente forma:

> "la fiabilidad es la seguridad y confianza de que un modelo actuará según lo previsto al enfrentarse a un problema determinado y funcionará correctamente con respecto a los objetivos y escenarios reales a los que se enfrenta, teniendo en cuenta que se puedan producir modificaciones en su parámetros o variables de entrada".

4.1.3. Causalidad

La siguiente propiedad a revisar es la Causalidad. Revisando la RAE, aparecen dos acepciones: una de ellas es la "Causa, origen, principio" mientras que la otra es: "Ley en virtud de la cual se producen efectos.". Se revisan ahora las definiciones dadas por los diferentes autores en la bibliografía específica para saber si se ajustan a lo descrito por la RAE.

La primera definición a revisar es la dada por Doshi-Velez et al. [Dos2017] que indica que la causalidad implica que el cambio predicho en la salida debido a una perturbación se producirá en el sistema real.

Al seguir profundizando en la bibliografía, Lipton [Lip2017] también habla de causalidad haciendo referencia a esta propiedad como aquella que permite a los investigadores inferir propiedades o generar hipótesis sobre el mundo natural. La tarea de inferir relaciones causales a partir de datos observacionales ha sido ampliamente estudiada, pero todos estos métodos tienden a basarse en fuertes suposiciones de conocimiento previo.

La última referencia revisada es la dada por Barredo et al. [Bar2020] en la que indica que la causalidad requiere un amplio marco de conocimiento previo para demostrar que los efectos observados son causales. Un modelo de aprendizaje automático descubre correlaciones entre los datos de los que aprende, por lo que podría no ser suficiente para desvelar una relación causa-efecto. Sin embargo, la causalidad implica correlación, por lo que un modelo de aprendizaje automático explicable podría validar los resultados proporcionados por las técnicas de inferencia de causalidad, o proporcionar una primera intuición de posibles relaciones causales dentro de los datos disponibles.

Una vez revisadas las diferentes acepciones de esta propiedad, se propone definir la causalidad como:

> "aquella propiedad que permite que los modelos obtengan correlaciones de los datos de los que aprenden, pudiendo así proporcionar una primera intuición de posibles relaciones eventuales dentro de estos datos".

4.1.4. Transferibilidad

Otra de las propiedades a revisar es la transferibilidad. Para la RAE su significado es que puede ser transferido o traspasado. Se revisan ahora las diferentes referencias encontradas en la bibliografía especializada.

La primera referencia se encuentra en Lipton [Lip2017]. En ella se indica que un modelo debe de tener la capacidad de transferir sus habilidades aprendidas a otros escenarios que no son familiares al igual que los humanos tienen la capacidad de poder generalizar cuando ocurren estas situaciones.

Otra de las definiciones encontradas es la dada por Barredo et al. [Bar2020] en la que se indica que la transferibilidad es la mera comprensión de las relaciones internas que tienen lugar dentro de un modelo que facilita la capacidad de un usuario para reutilizar este conocimiento en otro problema. Es por ello, que la transferibilidad

debería ser una de las propiedades de un modelo explicable, pero que se verifique esta propiedad en el modelo no implica que sea explicable.

Vistas estas definiciones, se propone definir la transferibilidad como:

> "la propiedad por la cual un usuario puede aplicar de forma generalizada el conocimiento aprendido mediante un modelo a modelos que se dan en situaciones distintas".

4.1.5. Informatividad

La siguiente propiedad a revisar es Informatividad. Al igual que ocurre con Transferibilidad, las dos referencias estudiadas, además de la definición de la RAE, serán las dadas por Lipton [Lip2017] y Barredo et al. [Bar2020].

Comenzando por la acepción encontrada en la RAE, esta nos indica que la informatividad es la comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada.

Para Lipton [Lip2017] la informatividad es la forma más obvia en la que un modelo transmite información a través de sus resultados, permitiendo, a través de algún procedimiento, transmitir información adicional al decisor humano.

Mientras que para Barredo et al. [Bar2020] es importante que los modelos explicables deben dar información sobre el problema que se aborda, de forma que se sea capaz de relacionar la decisión del usuario con la solución dada por el modelo y evitar caer así en errores en esta relación.

Vistas estas definiciones, se propone definir la informatividad como:

> "el proceso por el cual el modelo debe transmitir información tanto acerca de sus resultados como del problema que se aborda para poder entender y relacionar la decisión tomada."

4.1.6. Imparcialidad

El siguiente término a estudiar será el de imparcialidad. Según la RAE imparcialidad es la falta de designio anticipado o de prevención en favor o en contra de alguien o algo, que permite juzgar o proceder con rectitud. Revisamos ahora las diferentes definiciones encontradas en la bibliografía especializada.

La primera referencia bibliográfica encontrada es la dada por Lipton [Lip2017], que habla de la toma de decisiones justa y ética, *fair and ethical decision-making,* y como cada vez se está haciendo más presente la preocupación de la evaluación e interpretación de las decisiones producidas automáticamente para verificar que las mismas se ajustan a las normas éticas.

La siguiente referencia encontrada es la dada por Doshi-Velez et al. [Dos2019], que indica que la imparcialidad implica que los grupos protegidos (explícitos o implícitos) no son discriminados de alguna manera.

Por último, la siguiente referencia a estudiar es la dada por Barredo et al. [Bar2020] que indica que la imparcialidad es la capacidad de garantizar y alcanzar equidad en los modelos de aprendizaje automático, visto desde un punto vista social. Además, estos autores, destacan que un objetivo relacionado con XAI es el de destacar los sesgos en los datos a los que se expuso el modelo.

Teniendo en cuenta las distintas definiciones, se propone definir la imparcialidad como:

> “la propiedad por la cual los modelos de inteligencia artificial explicable deben garantizar una toma de decisiones justa y ética cuando se decide un resultado a partir de la salida del modelo, de la misma forma que los datos utilizados, tanto en el entrenamiento como durante la ejecución del modelo, no deben de estar perjudicados por ningún tipo de sesgo, ya sea en datos, algoritmos u otros”.

4.1.7. Interactividad

El último término a estudiar será el de interactividad. Al igual que en los anteriores, se comienza con la definición dada por la RAE para este término. Según la RAE, interactividad es una cualidad de interactivo y dicho de un programa informático, indica que permite una interacción, a modo de diálogo, entre la computadora y el usuario.

En Barredo et al. [Bar2020] se hace referencia a la interactividad indicando que uno de los objetivos de un modelo de aprendizaje automático y por ende de la IA, es que sea interactivo para el usuario. ¿Cómo se puede conseguir un modelo que sea interactivo para el usuario? Montavon et al. [Mon2018] indican que para ello es necesario que el modelo se realice bajo un dominio que sea interpretable para el tomador de decisiones.

De acuerdo con esto, se propone definir la interactividad como:

> "la forma en la que un modelo es capaz de transmitir de forma simple y comunicativa su resultado, ya sea a través de visualizaciones gráficas, lenguaje natural, etc. Para ello, ha de hacerlo bajo el dominio interpretable del usuario".

4.2. Resumen de las propiedades y objetivos de un modelo XAI

Al igual que en el capítulo anterior, en este capítulo también se ha desarrollado un resumen final de todas las propiedades y objetivos estudiados, la definición propuesta así como las referencias a la bibliografía especializada que se ha analizado.

Concepto	Definición	Referencias
Privacidad	Para que un modelo cumpla la propiedad de privacidad, este debe de ser capaz de proteger toda la información sensible que se maneje en cada una de sus etapas, independientemente de que el modelo sea entrenado por nosotros mismos o por terceros, siendo particularmente riguroso en las etapas de explicación del modelo para evitar exponer la privacidad de los datos sensibles al intentar hacer más explicable el modelo.	RAE [Bar2020] [Dos2017] [Lip2017]
Fiabilidad	Es la seguridad y confianza de que un modelo actuará según lo previsto al enfrentarse a un problema determinado y funcionará correctamente con respecto a los objetivos y escenarios reales a los que se enfrenta, teniendo en cuenta que se puedan producir modificaciones en su parámetros o variables de entrada.	RAE [Bar2020] [Dos2017] [Lip2017]
Causalidad	Es aquella propiedad que permite obtener correlaciones de los datos que aprende de forma que pueda proporcionar una primera intuición de posibles relaciones eventuales dentro de estos datos.	RAE [Bar2020] [Dos2017] [Lip2017]
Transferibilidad	Es la propiedad por la cual un usuario puede aplicar de forma generalizada el conocimiento aprendido mediante un modelo a modelos que se dan en situaciones distintas.	RAE [Bar2020] [Lip2017]
Informatividad	Es el proceso por el cual el modelo debe transmitir información tanto acerca de sus resultados como del problema que se aborda para poder entender y relacionar la decisión tomada.	RAE [Lip2017] [Bar2020]
Imparcialidad	Propiedad por la cual los modelos de inteligencia artificial explicable deben garantizar una toma de decisiones justa y ética cuando se decide un resultado a partir de la salida del modelo, de la misma forma que los datos utilizados, tanto en el entrenamiento como durante la ejecución del modelo, no deben de estar perjudicados por ningún tipo de sesgo, ya sea en datos, algoritmos u otros	RAE [Bar2020] [Dos2017] [Lip2017]
Interactividad	Forma en la que un modelo es capaz de transmitir de forma simple y comunicativa su resultado, ya sea a través de visualizaciones gráficas, lenguaje natural, etc. Para ello, ha de hacerlo bajo el dominio interpretable del usuario.	RAE [Bar2020] [Mon2018]

Tabla 2. Resumen de las propiedades y objetivos de un modelo XAI

5. Métricas en Inteligencia Artificial Explicable

Una vez revisados los diferentes conceptos, propiedades y objetivos que deben de estar presentes en un modelo XAI, se revisan distintas métricas que permitan cuantificar el nivel (o grado) de explicabilidad de un modelo.

Se puede comprobar que diferentes autores hablan de la medición de la explicabilidad de un modelo de distintas formas. Por ejemplo, en el caso de Hoffman et al. [Hof2019] y Mosheni et al. [Mos2020], realizan esa medición a partir de entrevistas y/o encuestas al tomador de decisiones. Por otra parte, hay otros autores que hablan de la medición de la explicabilidad a partir de las propiedades del mismo, como es el caso de Himabindu et al. [Him2016], que describe propuestas para la medición de la explicabilidad en modelos de IA basados en reglas.

En definitiva se puede considerar que la explicabilidad se mide o bien de forma subjetiva basada en la percepción del modelo por parte del tomador de decisiones (recogida a través de encuestas) o de forma objetiva en base a la cuantificación de las distintas propiedades y objetivos de un modelo XAI.

Las medidas objetivas propuestas hasta el momento dependen del tipo de esquema de representación del conocimiento. Los modelos basados en reglas representan uno de los paradigmas de IA más utilizados cuando se requiere conocimiento explicable. Por ello en este capítulo se realizará una revisión de las métricas de tipo objetivas para un modelo de IA basado en reglas.

5.1. Métricas para modelo de Inteligencia Artificial basados en reglas

Son distintos autores los que hablan de métricas en modelos de IA basados en reglas. En esta sección se lleva a cabo un repaso por los diferentes autores y propuestas de medición de interpretabilidad que presentan.

Una de las indicaciones que dan He et al. [He2020] para ver la interpretabilidad de un modelo está relacionada con el número de reglas y el número de condiciones de las mismas. Estos autores indican que para que sea interpretable por un ser

humano, el número de reglas ha de ser el más bajo posible y se apoyan en lo indicado por Miller [Mil1956] para concluir que el número de condicionantes de estas reglas debe de estar entre 5-9, ya que es el límite que un humano es capaz de manejar.

Para reducir el número de reglas, Mikut et al. [Mik2005] propone incluir en el propio modelo un algoritmo de selección de reglas interactivo que cuantifique la interpretabilidad a través del número de condiciones en el antecedente de las reglas. Para ello, utiliza una medida de claridad (*clearness*) de la regla basada en la relevancia estadística de la regla y el nivel de generalización, que indica que cuanto mayor sea el resultado obtenido de esta medida, menor es el número de reglas utilizado y mayor por tanto su interpretabilidad. Esta medida también es utilizada por Gacto et al. [Gac2011] y He et al. [He2020].

Otra de las medidas que trata de cuantificar la relevancia de los factores incluidos en el antecedente de las reglas es la propuesta por [Ped2003] y que también ha sido recogida por He et al. [He2020]. Se trata de la consistencia del conjunto de reglas. Para ellos, la consistencia en las reglas se refiere a la diferencia de conclusiones que se obtienen con antecedentes similares, o lo que es lo mismo, mientras menor sea el número de ejemplos negativos con los que trabajamos en las reglas, más consistentes serán las mismas y mayor interpretabilidad tendrá el modelo.

Otros de los autores que hablan de la cuantificación de la interpretabilidad en un modelo de IA basado en reglas son Alonso et al. [Alo2008], en cuyo trabajo proponen la obtención del índice de interpretabilidad a partir de las siguientes características:

- número total de reglas, (c1)
- número total de premisas, (c2)
- número de reglas que utilizan una entrada, (c3)
- número de reglas que utilizan dos entradas, (c4)
- número de reglas que utilizan tres o más entradas y (c5)
- número de etiquetas definidas por variable. (c6)

De esta forma, se puede calcular el índice de interpretabilidad de un modelo a partir de la conexión de estas características como se observa en la ilustración 4.

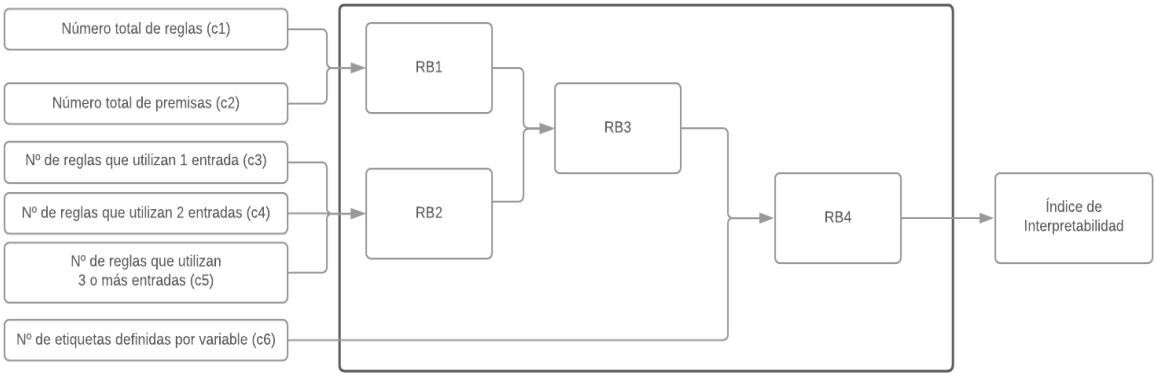

Ilustración 4. Explicación del cálculo del índice de Interpretabilidad. Fuente: [Alo2008]

Para obtener el índice de interpretabilidad como se puede observar en la figura 4, primero hay que conocer cómo calcular los diferentes valores intermedios (RB1, RB2, RB3 y RB4) que nos hacen llegar a ellos. Estos valores siempre oscilarán entre 0 y 1 y se calculan de la siguiente forma:

- RB1: realiza una estimación de la dimensión de la base de las reglas, utilizando c1 y c2 como entrada. Aunque c1 y c2 utilizan información similar tienen un significado diferente, mientras mayor sea el número de reglas y premisas menor será por tanto la interpretabilidad.
- RB2: evalúa la complejidad de las reglas a partir del número de entradas de las reglas, utilizando c3, c4 y c5 como entrada. La complejidad de este valor vendrá dada por el número de reglas del tipo c5, mientras más alto sea este valor con respecto a c3 y c4 más complejo y por tanto menos interpretable se volverá nuestro modelo.
- RB3: utiliza la dimensión de la base de las reglas (RB1) junto con la complejidad (RB2) para obtener el índice de interpretabilidad de las variables. Cuanto mayor sea la dimensión y/o complejidad de las reglas menor será la interpretabilidad del modelo.
- RB4: integra la interpretabilidad de las variables (RB3) junto con el número de etiquetas definidas por cada variable (c6). Si el número medio de etiquetas definidas por la entrada es bajo, cuanto mayor sea la

interpretabilidad de la base de reglas, mayor será el índice de interpretabilidad. Sin embargo, si el número medio de etiquetas definidas por la entrada es alto, el índice de interpretabilidad reproduce la interpretabilidad de la base de reglas pero proporcionalmente reducida.

Por último, se revisan las medidas propuestas por Lakkaraju et al. [Lak2016] para cuantificar la interpretabilidad de un modelo basado en reglas. Estos autores proponen cinco medidas: fracción de solapamiento, fracción de no cobertura, media del tamaño de las reglas, número de reglas y fracción de clase.

- Fracción de solapamiento: capta el grado de solapamiento entre cada par de reglas de un conjunto de decisiones. Mientras más pequeños sean estos valores, más interpretable es el modelo.
- Fracción de no cobertura: mide la fracción de datos que no están cubiertos por ninguna regla del conjunto. Su valor va de 0 a 1, donde 0 significa que todos los puntos están cubiertos e incrementa su valor hasta llegar a 1, que significa que ningún punto está cubierto. Mientras menor sea su valor, más interpretable será el modelo, ya que estarán cubiertos todos los puntos minimizando así el número de reglas utilizadas.
- Media del tamaño de las reglas: capta el número medio de predicados que un lector humano debe analizar para entender una regla en un conjunto de decisiones. Mientras menor sea el tamaño, más interpretable será para el ser humano.
- Número de reglas: es el número de reglas del conjunto a medir. Al igual que ocurre con el tamaño de las reglas, mientras menor sea este valor, mayor será la interpretabilidad.
- Fracción de clases: mide qué fracción de las etiquetas de clase en los datos son predichas por al menos una regla en un conjunto de decisiones. Esta métrica tiene un límite inferior a 0 que corresponde a la situación en la que el conjunto de reglas no describe información sobre ninguna clase, lo que podría significar que el conjunto está vacío. El valor máximo que puede tomar esta métrica es 1, que corresponde al caso en el que cada clase está descrita por alguna regla del conjunto. Que todas las clases

estén descritas dentro del conjunto de reglas permite una mayor interpretabilidad del modelo.

Una vez revisadas todas las medidas propuestas por los diferentes autores, se concluye que la medición de un modelo de IA basado en reglas puede cuantificar su interpretabilidad mediante:

- El número de reglas,
- el tamaño de las reglas o número de condiciones en el antecedente, y
- la consistencia del modelo

De forma que, mientras menor sea el número de reglas utilizadas, el número de condiciones en el antecedente que cubren todas las necesidades del modelo y menor sea la diferencia de conclusiones que se obtienen con antecedentes similares en el conjunto de reglas, mayor será la interpretabilidad de este modelo para un ser humano.

5.2. Resumen de las métricas para modelos de Inteligencia Artificial basados en reglas

Al igual que en los capítulos anteriores, en este capítulo también se ha desarrollado un resumen final de las métricas para modelos de IA basados en reglas que se han propuesto, así como su influencia en la interpretabilidad de un modelo como las referencias a la bibliografía especializada que se ha analizado.

Medida	Influencia en la interpretabilidad del modelo	Referencias
Número de reglas	El número de reglas debe de ser pequeño para que el modelo pueda ser más interpretable.	[He2020] [Lak2016] [Alo2008]
Tamaño de reglas	Mientras más pequeño sea el tamaño de cada una de las reglas que forman el conjunto del modelo, mayor será la interpretabilidad del mismo.	[He2020] [Mil1956] [Lak2016] [Alo2008]
Consistencia	Mientras menor sea la diferencia de conclusiones que se obtienen de un conjunto de reglas con antecedentes similares, mayor será la interpretabilidad del modelo.	[Ped2003] [He2020]

Tabla 3. Resumen de las métricas aplicables a un modelo de IA basado en reglas

6. Conclusiones y líneas de trabajo futuras

En esta sección se exponen las conclusiones obtenidas en este trabajo fin de máster. Además de estas conclusiones se señalan algunos trabajos futuros que se podrían realizar siguiendo las líneas de actuación trazadas en este trabajo.

Indicar que durante este estudio se han revisado en torno a 200 referencias bibliográficas en las que se menciona el concepto de XAI o algunas de sus propiedades, objetivos y/o métricas. Las referencias que aparecen en el apartado final de bibliografía son las que se han citado en el estudio ya que han sido las que nos han aportado una visión completa de la terminología.

6.1. Conclusiones

En la bibliografía especializada en XAI no existe un consenso respecto a la denominación y definición de los conceptos utilizados en el área. En este trabajo se describen parte de los resultados del análisis de la bibliografía existente con el objetivo de extraer información y fomentar el debate en el área que permita llegar a una terminología unificada. Debido a la gran importancia que día a día está cobrando XAI, se considera importante realizar una unificación de esta terminología para poder sentar las bases en las que seguir trabajando.

En este estudio, se ha realizado un análisis exhaustivo de las referencias encontradas en la bibliografía especializada sobre XAI, que ha permitido diferenciar y unificar -según el caso- conceptos, proponer definiciones integradoras y avanzar en el conocimiento sobre propiedades de un modelo XAI que serán objetivos a considerar en su diseño y/o evaluación.

Además de avanzar en toda la terminología XAI, también se ha realizado una revisión de cómo poder cuantificar la interpretabilidad en un modelo de IA, centrándonos en este estudio en los modelos de IA basados en reglas.

Cabe indicar que parte del estudio realizado para este trabajo fin de máster se ha presentado como artículo para la XIX Conferencia de la Asociación Española de Inteligencia Artificial (CAEPIA), el cual ha sido aceptado para su publicación y presentación en la misma [Esc2021].

6.2. Líneas de trabajo futuras

Es mucho lo que queda por hacer en esta rama de la IA, las líneas de trabajo futuras que se nos abren con este trabajo inicial se dividen claramente en dos.

La primera de ella es seguir trabajando para que la unificación de terminología propuesta en este estudio a partir del debate en foros científicos, y mejorando a la vez que ampliando la misma para que pueda servir como base en futuros estudios consiguiendo incluso generar una taxonomía propia de XAI basada en las diferentes propiedades y objetivos que lo rodean y que la misma sea aceptada por la comunidad. La segunda línea de trabajo futuro tiene que ver con la medición y cuantificación de la interpretabilidad de un modelo, aunque en este trabajo se ha hablado de modelos de IA basados en reglas son muchos más los modelos de IA existentes y pendientes de analizar, es más, existen modelos IA, especialmente los modelos de caja negra, como redes neuronales, aprendizaje profundo, máquinas de soporte vectorial, ensembles, etc., en los cuales no se podrá realizar una medición de la interpretabilidad o en el caso de poderse, será mucho más compleja de obtener que en otros tipos como pueden ser los modelos de IA basados en arboles.

Bibliografía

[Mil1956] G.A. Miller; The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev., 63(2) (1956), 81–97.

[Mic1983] R.S. Michalski; A theory and methodology of inductive learning, in: Machine learning, Springer, 1983, pp. 83–13

[Ped2003] W. Pedrycz; Expressing Relevance Interpretability and Accuracy of Rule-Based Systems, Springer Berlin Heidelberg, pp. 546–567.

[Len2004] M. van Lent, W. Fisher, and M. Mancuso; An explainable artificial intelligence system for small-unit tactical behavior, in Proc. 16th Conf. Innov. Appl. Artif. Intell., 2004, pp. 900907.

[Mik2005] R. Mikut, J. Jäkel, L. Gröll, Interpretability issues in data-based learning of fuzzy systems, Fuzzy Sets Syst. 150 (2) (2005) 179–197.

[Alo2008] J.M. Alonso, L. Magdalena, S. Guillaume; HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism, Int. J. Intell. Syst. 23 (7) (2008) 761–794.

[Gac2011] M.J. Gacto, R. Alcalá, F. Herrera; Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Information Sciences 181 (2011), 4340-4360

[Fre2014] A. A. Freitas; Comprehensible classification models: A position paper. ACM SIGKDD Explor. Newslett. 15, 1 (2014), 1–10.

[Gun2016] David Gunning; DARPA, Explainable Artificial Intelligence (XAI), IJCAI-16 DLAI WS, 2016

[Lak2016] H. Lakkaraju, S. H Bach, and J. Leskovec; Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1675–1684, San Francisco, California, USA, 2016. ACM. doi: 10.1145/2939672.2939874.

[Dos2017] Finale Doshi-Velez and Been Kim; Towards a rigorous science of interpretable machine learning. 2017. [Online] Disponible: arXiv:1702.08608v2.

[Lip2017] Z. C. Lipton; The mythos of model interpretability, CoRR, vol. abs/1606.03490, 2017. [Online]. Disponible: https://arxiv.org/abs/1606.03490. Accessed on: Nov. 21, 2018.

[Ada2018] A. Adadi, M. Berrada; Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), 2018

[Bar2018] S. Barocas, S. Friedler, M. Hardt, J. Kroll, S. Venka-Tasubramanian, and H. Wallach; The FAT-ML Workshop Series on Fairness, Accountability and Transparency in Machine Learning. Accessed: Jun. 6, 2018.

[Fer2019] A. Fernandez , F. Herrera , O. Cordon , M. Jose del Jesus , F. Marcelloni; Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEEE Computational Intelligence Magazine 14 (1) (2019) 69–81

[FIC2018] FICO Community. (2018); Explainable Machine Learning Challenge. Accessed: Jun. 6, 2018.

[Gui2018] R. Guidotti , A. Monreale , S. Ruggieri , F. Turini , F. Giannotti , D. Pedreschi; A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018) 93:1–93:42

[Gun2018] D. Gunning; Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA). Accessed: Jun. 6, 2018.

[Mon2018] G. Montavon, W. Samek, K.-R. Müller; Methods for interpreting and understanding deep neural networks, Digital Signal Processing 73 (2018) 1–15, [Online] Disponible: doi: 10.1016/j.dsp.2017.10.011

[Gun2019] D. Gunning, D. W. Aha; DARPA's Explainable Artificial Intelligence Program, Summer 2019

[Hof2019] R. R Hoffman, S. T. Mueeler, G. Klein, J. Litman; Metrics for Explainable AI: Challenges and Prospects, 2019. [Online] Disponible: arXiv:1812.04608v2

[Mur2019] W. James Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl y B. Yu, ``Definitions, methods, and applications in interpretable machine learning". Proceedings of the National Academy of Sciences (44) (2019) 22071-22080, http://dx.doi.org/10.1073/pnas.1900654116

[Bar2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera; Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 2020

[Dye2020] M. Dye, C. Ekanadham, A. Saluja, A. Rastogi; Supporting content decision makers with machine learning, 2020, [Online]. Disponible: https://netflixtechblog.com/supporting-content-decision-makers-with-machine-learning-995b7b76006f. Accessed on: May. 20, 2021

[He2020] C. He, M. Ma, P. Wanga. Extract interpretability-accuracy balanced rules from artificial neural networks: A review. Neurocomputing 387 (2020) 346-358.

[Mos2020] S. Mosheni, N. Zarei y E.D. Ragan; A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI System, April. 2020. [Online] Disponible: arXiv:1811.11839v4

[Esc2021] A. Escobar, P. González, M.J. del Jesus; Revisión y análisis de conceptos en Inteligencia Artificial Explicable. Una aproximación a la unificación de la terminología, XIX Conferencia de la Asociación Española de Inteligencia Artificial (CAEPIA) (2021) (Pendiente publicación)

[Ism2021] A.M. Ismael, A. Sengür; Deep learning approaches for COVID-19 detection based on chest X-ray images, Expert Systems with Applications (164) 2021, [Online]. Disponible: https://doi.org/10.1016/j.eswa.2020.114054