Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial

Álvaro Navarro[1], Ana Lavalle[1], Alejandro Maté[1], and Juan Trujillo[1]

Lucentia Research Group, Departamento de Lenguajes y Sistemas Informáticos (DLSI), Universidad of Alicante,
Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690, España
alvaro.nl@dlsi.ua.es

Resumen Pese a la relevancia de los sistemas de Inteligencia Artificial (IA) y los modelos de aprendizaje automático (ML) en la sociedad, dichos sistemas y modelos a menudo están limitados por la opacidad en su toma de decisiones. Este punto es clave, pues la falta de interpretabilidad puede suponer que se tomen decisiones injustas de manera oculta, lo que impediría tomar acciones correctivas para solucionar el problema. Aunque diferentes trabajos han afrontado el desafío de la opacidad desde un punto de vista explicable, hay una carencia de propuestas que traten de explotar la información de los datos para este fin. En este contexto, y para ayudar al experto en ML en el análisis y la toma de decisiones, presentamos un proceso basado en un algoritmo jerárquico, denominado Árbol de Detección de Equidad (ADE), el cual recorre recursivamente los datos para crear un árbol de análisis (AAE). Combinado con técnicas de explicabilidad, unimos las características de los grupos a las decisiones del modelo, proporcionando a través del AAE información que puede mejorar la confianza en los resultados y propiciar una mejor comprensión del proceso de toma de decisiones del modelo. Las principales contribuciones de este trabajo son: (i) definimos métricas de equidad que tienen en cuenta conjuntos de datos reducidos y/o desbalanceados; (ii) analizamos automáticamente el conjunto de datos explotando los pesos extraídos del modelo; e (iii) identificamos grupos cuyo trato ha sido potencialmente injusto. Para demostrar la aplicabilidad de nuestra propuesta, analizamos su efectividad en cuatro dominios distintos.

Keywords: Equidad · Inteligencia Artificial · Inteligencia Artificial Explicable · Análisis de datos

1. Introducción

Dada la influencia que han adquirido los sistemas de Inteligencia Artificial (IA) en nuestra sociedad, su estudio es sin duda una de las mayores necesidades de la actualidad. Debido a ello, hay una amplia variedad de dominios donde los sistemas de IA trabajan en nuestra vida cotidiana, incluso en dominios críticos como la predicción de inundaciones [24]. Tanto gobiernos como grandes compañías han delegado responsabilidades en estos sistemas, en algunos casos con

problemas derivados de decisiones injustas que han tomado. La raíz de tales problemas es la diferencia en el trato hacia determinados grupos o individuos, derivados desde el modelo priorizando la atención en algunos atributos específicos de los datos. A menudo estos atributos son sensibles (como el sexo o la raza) y se identifican grupos que pueden ser categorizados alrededor de tales atributos sensibles (en adelante, se les denominará atributos protegidos y grupos protegidos). Por ejemplo, ha habido problemas discriminando la raza (en [20], un software fue usado por jueces para decidir si una persona debía seguir en prisión o no, y los resultados demostraron un sesgo hacia algunas razas). Éste es sólo un ejemplo de un comportamiento discriminatorio hacia un grupo protegido, demostrando la relevancia de entender y evaluar la equidad de un sistema de IA antes de que dicho sistema llegue a su aplicación en la sociedad [2].

De esta forma, que los sistemas de IA tomen decisiones justas se ha convertido en algo cada vez más solicitado [12]. Para ello, emergió un concepto llamado IA Ética (también IA responsable e IA confiable) [5]. La IA ética trata de evitar que los modelos de aprendizaje automático (ML), los cuales son subsistemas de IA, tomen decisiones injustas, paliando así los problemas mencionados anteriormente. Para afrontar dichos problemas o desafíos, otro concepto fundamental es el de IA explicable (XAI). XAI trata de arrojar luz sobre los modelos que se comportan como caja negra, descifrando las razones por las que dichos modelos toman sus decisiones [2]. Este punto es fundamental, ya que una forma de evaluar la equidad de un sistema de IA es entender su proceso de toma de decisiones. En este contexto, hay diferentes técnicas de explicabilidad que proporcionan información a los desarrolladores o analistas (en adelante denominados expertos en ML). Esta información les ayuda a revisar y analizar cómo el modelo está tomando sus decisiones. Debido a esto, entender el proceso de toma de decisiones del modelo de ML es crucial para desarrollar correctamente los sistemas de IA y, posteriormente, aplicar dicho sistema en la sociedad.

En consecuencia, proponemos unir los conceptos de IA ética y XAI, e intentamos afrontar el desafío de las decisiones injustas enfocándonos en explicar las decisiones teniendo en cuenta tanto los datos como la evaluación de la equidad. Con este fin, presentamos un proceso basado en un novedoso algoritmo jerárquico, llamado Árbol de Detección de Equidad (ADE), el cual proporciona información para trabajar con conjuntos de datos y entender el comportamiento de la equidad en éstos. El algoritmo recorre recursivamente los subgrupos (categorías) de las variables (atributo, columna o característica) de los datos y calcula su equidad. El ADE ejecuta su búsqueda en base a los pesos de las variables del modelo de ML en la toma de decisiones, obtenidos mediante una técnica de explicabilidad de extracción de características [2]. Después, el ADE ejecuta un análisis completo en los diferentes niveles. Estos niveles corresponden a las variables más importantes de acuerdo a los pesos anteriormente extraídos. Dado que cada nivel no depende del resto, este algoritmo permite hacer un análisis no sólo entre grupos sino también intragrupo. Una vez el ADE se ha ejecutado, un Árbol de Análisis de Equidad (AAE) es generado como salida, dónde la equidad y la información de los datos es almacenada y está lista para ser presentada a

los expertos en ML. Debido a que el ADE construye el AAE a través de los niveles explicados anteriormente, dicha información se almacena y presenta de forma ordenada y jerárquica. Gracias a nuestra propuesta, los expertos en ML son capaces de analizar el comportamiento de la equidad en los datos. Finalmente, para poder trabajar correctamente con conjuntos de datos reducidos y/o desbalanceados, definimos y hacemos uso de dos nuevas métricas de equidad en la IA: el Ratio de Verdaderos Ponderado (*RVP*) y el Ratio de Falsos Ponderado (*RFP*).

El resto del artículo se estructura de la siguiente manera. Nuestra propuesta se presenta en la Sección 2. En la Sección 3 se expone el caso de estudio dónde nuestra propuesta se ha aplicado. El estado de la cuestión se presenta en la Sección 4. Finalmente, las conclusiones y el trabajo futuro se presentan en la Sección 5.

2. Proceso para la evaluación de la equidad

Para abordar el problema de los modelos de ML que pueden llegar a tomar decisiones injustas, presentamos un proceso para evaluar la equidad en dichos modelos. Como muestra la Figura 1, nuestro proceso se basa en: (i) obtener los datos de entrada (representados en color azul) para la correcta ejecución de (ii) el algoritmo Árbol de Detección de Equidad (ADE, representado en color rojo) que recorrerá exhaustivamente los datos y así (iii) generar una salida estructurada en el Árbol de Análisis de Equidad (AAE, representado en color verde) para poder tener más información de cómo se comportan los modelos de ML y detectar problemas que podrían haber pasado inadvertidos. A continuación, se detallan los pasos del proceso presentado.

2.1. Entrada

Datos:

Primero, se debe definir el conjunto de datos con el que se va a trabajar. Para poder aplicar el algoritmo ADE, el dataset debe ser categórico, donde las filas

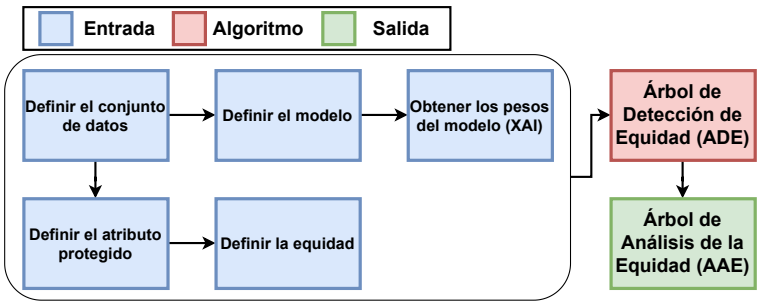

Figura 1. Proceso propuesto para evaluar la equidad.

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

representan las instancias y cada una de éstas tiene un valor (grupo o categoría) para el atributo (variable, columna o característica) específico. Este paso es esencial ya que dicho algoritmo busca recursivamente **subgrupos** o **categorías** en las variables de los datos. Por consiguiente, las variables no categóricas deben ser categorizadas para el correcto funcionamiento del algoritmo y del proceso (en el caso de valores continuos, esto se puede conseguir mediante un proceso de discretización).

Modelo:

Tras obtener el conjunto de datos categórico, se debe **entrenar el modelo de ML**. Dado que los datos de los se parte están etiquetados, el entrenamiento del modelo consiste un proceso de aprendizaje supervisado. Las predicciones del modelo serán almacenadas en un estructura de datos (*dataframe*) que contendrá tanto (i) la clase predicha como (ii) la clase real. Sin embargo, la categorización y discretización mencionadas anteriormente pueden afectar a cómo aprende el modelo de ML. Dada esta situación, el modelo específico debe ser analizado y, si un experto en ML involucrado en el proceso concluye que hay limitaciones significativas dado el tratamiento de datos realizado, se deben mitigar dichas limitaciones antes de continuar con los pasos incluidos en el proceso propuesto. Una vez definido el modelo, se extraen los **pesos** del modelo. Para ello, se debe aplicar una técnica de explicabilidad post-hoc enfocada en la extracción de características [2]. Además, aunque a menudo estas técnicas están enfocadas en interpretar instancias individuales de un objeto (ámbito local del modelo), también se pueden utilizar para analizar el modelo completo (ámbito global). La información acerca de los pesos del modelo permitirá analizar qué importancia le da el modelo a cada variable, lo que es fundamental para entender el proceso de toma de decisiones de dicho modelo.

Equidad:

El siguiente paso es observar si hay o no un **atributo protegido** en los datos. En este contexto, la literatura establece distintos tipos de atributos protegidos (también llamados sensibles) [17]. De esta forma, un atributo protegido no sólo puede aparecer como un grupo general que diferencia un aspecto que la sociedad entiende no se debe discriminar (raza, sexo, etc.), sino también a través de variables representantes (*proxy*). Por ejemplo, si analizamos la variable barrio, se debe estudiar la correlación entre dicha variable y la tasa de personas de cada raza de ese barrio. Contrario a esto, también se debe estudiar si, para un caso específico, se debe considerar como correcto o adecuado hacer una pseudo-discriminación como, en el caso de estudiar el cáncer de mama, tener en cuenta el sexo del o de la paciente en cuestión. De esta forma, aunque pueda haber o no un atributo protegido en el conjunto de datos (con independencia de cómo de visible o escondido se encuentre éste), poder estudiar exhaustivamente cómo cada variable opera es otra de las bondades del proceso -y el algoritmo- propuesto.

Una vez sabemos si hay o no atributo protegido y, si lo hay, dependiendo de cuál es, se puede definir qué **métrica** (o métricas) se utilizará para medir la **equidad**. Gracias a ello, es posible evaluar un sistema de IA y evitar que

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

éste tome decisiones injustas. De esta forma, hay diferentes tipos de equidad que se deben considerar a la hora de evaluar la equidad del sistema de IA, en concreto: (i) equidad de los individuos [10]: trata de, dados individuos similares en el espacio de datos, observar si las predicciones son similares y gracias a ello garantizar un trato justo; (ii) equidad de grupo [10]: su objetivo es evaluar cómo distintos grupos son tratados e intentar evitar un comportamiento del modelo discriminatorio; y (iii) equidad de los subgrupos [15]: su objetivo es capturar comportamientos no deseados tanto en grupos como en individuos. Para medir cada tipo de equidad, hay diferentes métricas de equidad propuestas en la literatura. Algunas de estas métricas están basadas en evaluar los ratios dada la matriz de confusión de las predicciones del modelo, como la igualdad de oportunidades (*Equal Opportunity* [13]) que se centra en el ratio de positivos (*True Positive Rate*, *TPR*), la igualdad predictiva (*Predictive Equality* [5]) que tiene el foco en evaluar el ratio de negativos (*False Positive Rate*, *FPR*), o las probabilidades igualadas (*Equalized Odds* [13]) que tiene en cuenta tanto el ratio de positivos (*TPR*) como el de negativos (*FPR*).

Además, cabe destacar que el proceso propuesto es capaz de trabajar con cualquier atributo protegido y cualquier métrica para evaluar equidad, pues el algoritmo ADE está habilitado para ello. Sin embargo, la elección por parte del analista o experto en ML es importante para poder evaluar correctamente el escenario específico.

2.2. Algoritmo: Árbol de Detección de Equidad (ADE)

La piedra angular del proceso propuesto es el algoritmo Árbol de Detección de Equidad (ADE) que busca exhaustivamente información que puede ser útil en el análisis del comportamiento del modelo de ML. Además, su naturaleza recursiva permite buscar rápidamente y luego obtener de manera jerárquica y ordenada toda la información recogida.

Entrando en detalle, el algoritmo ADE recibe como entrada: (i) los datos categóricos, (ii) una matriz donde los pesos de los atributos están almacenados, (iii) las columnas (variables) para buscar en las siguientes iteraciones, (iv) los valores reales de la clase a predecir, (v) los valores predichos de la clase a predecir, (vi) el mínimo número de filas para ejecutar el algoritmo ADE, (vii) la profundidad máxima para seguir ejecutando el algoritmo, (viii) el atributo protegido, y (ix) un árbol vacío dónde se irá almacenado la salida del ADE: el Árbol de Análisis de Equidad (AAE).

De esta forma, el algoritmo ADE comienza verificando los casos base. Para ello, se han establecido **tres casos base** distintos: (i) si no hay suficientes atributos (columnas) para seguir con la búsqueda o (ii) si no hay suficiente número de filas para operar de manera confiable, el algoritmo ADE devuelve un árbol vacío. Si, por otra parte, (iii) la profundidad se ha alcanzado, se devuelve el árbol del nivel actual. Si estos casos base no se alcanzan, el algoritmo ADE comienza su **búsqueda recursiva**. Para ello, inicializa el AAE donde irá almacenando la información por niveles con una estructura jerárquica de nodos de un árbol (a

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

nivel de programación, en Python se traduce a un diccionario clave-valor) empezando por el nivel actual donde considera las posibilidades para la variable actual (primera en buscar). En este proceso de almacenamiento de información, se realiza el cálculo de la métrica de equidad definida para, posteriormente, guardar el valor resultante en el nivel correspondiente. Después, se deben actualizar los parámetros para saber cuáles son ahora las variables de más peso y las siguientes a buscar. Una vez hecho ésto, el algoritmo ADE hace la llamada recursiva para ir construyendo la totalidad del árbol. La principal bondad que le aporta la recursividad a este algoritmo de búsqueda exhaustiva es que le permite ir almacenando la información jerárquicamente para luego obtener una salida correctamente estructurada en el AAE.

2.3. Salida: Árbol de Análisis de Equidad (AAE)

El paso final del proceso propuesto es el Árbol de Análisis de Equidad (AAE), que es generado como salida del ADE. La Figura 2 muestra cómo el AAE se construye, donde para cada nivel (llamada recursiva) hay diferentes pasos para ejecutar. Como muestra la Figura, los pasos relativos a los pesos se muestran en color verde, los de filas de datos en color azul, los de la equidad en color rojo, las combinaciones y sus grupos en color amarillo, y las métricas de ML de la matriz de confusión y sus ratios en color violeta. En concreto, esos pasos son los siguientes:

1. El nodo padre es seleccionado por ser el de mayor peso.
2. Para dicho atributo, se muestra cada combinación posible dentro de sus categorías, presentando para cada combinación: el peso del atributo, el valor de la métrica de equidad y las siguientes columnas a iterar en función de los pesos (junto con éstos).

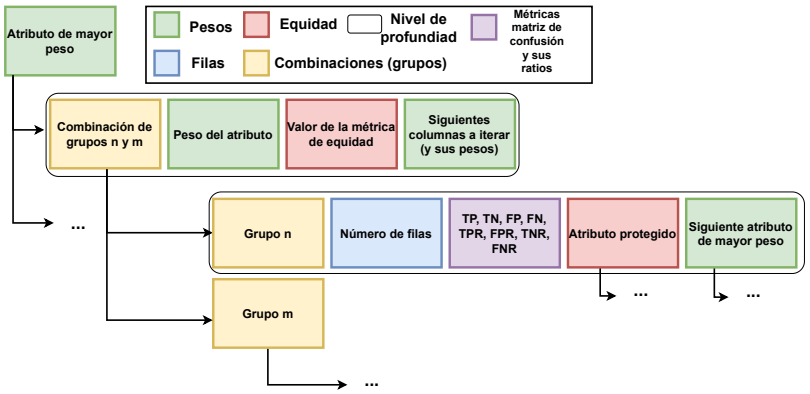

Figura 2. Proceso para generar el Árbol de Análisis de Equidad (AAE)

3. Dentro de cada combinación hay dos grupos. Para cada grupo se muestra su número de filas, sus métricas extraídas del rendimiento del modelo de ML[1] (matriz de confusión del modelo) junto con sus ratios[2].
4. Si hay atributo protegido, se muestra esta misma información para las categorías de dicho atributo.
5. Finalmente, si se debe seguir iterando, el proceso se repite con el siguiente atributo de mayor peso.

Una vez esta información ha sido volcada en el AAE, los expertos en ML tendrán una capacidad mayor para poder analizar y tomar decisiones sobre los sistemas de IA (o modelos de ML) y así observar problemas que pudieran haber pasado inadvertidos. Más en concreto, se podrá observar: (i) cuáles son los atributos de más peso (dónde se aplica la técnica XAI de extracción de características); (ii) las características de la distribución de los datos; y (iii) la equidad calculada.

3. Caso de estudio: Adult (Census Income)

A continuación, se detalla cómo se ha evaluado nuestra propuesta, especificando los pasos del proceso propuesto (Figura 1).

3.1. Entrada

Datos:

El conjunto de datos de Adult (Census Income) [9] está relacionado con la estimación de los ingresos de personas adultas, es decir, evalúa las variables para predecir cuántos ingresos se tendrán en función de distintas variables. Sin embargo, este conjunto de datos ha sido ampliamente estudiado por haber tomado decisiones injustas hacia los atributos protegidos que tiene, infraestimando al colectivo femenino. Analizándolo en detalle, contiene información sobre 32.561 personas (filas) y 14 variables de entrada para predecir la variable de salida "Income" que representa el total de ingresos de la persona (fila) en cuestión.

Modelo:

Para entrenar al modelo, se ha seleccionado el clasificador XGBoost [7], que es uno de los mejores y más conocidos modelos de ML por su nivel de rendimiento. En concreto, hemos utilizado un tamaño para entrenar del 75 % de la muestra, y el resto 25 % lo hemos usado para probar el modelo. También hemos usado una función de evaluación logloss (negative log-likelihood) y un tamaño de paso (eta: η) de 0.3. En la Tabla 1 podemos observar el rendimiento del modelo para el conjunto de datos de Adult [9]. Para poder presentar una evaluación completa,

[1] Verdaderos Positivos (TP), Falsos Positivos (FP), Verdaderos Negativos (TN) y Falsos Negativos (FN).

[2] Ratio de Verdaderos Positivos ($TPR = \frac{TP}{TP+FN}$), Ratio de Verdaderos Negativos ($TNR = \frac{TN}{TN+FP}$), Ratio de Falsos Negativos ($FNR = \frac{FN}{FN+TP}$), y Ratio de Falsos Positivos ($FPR = \frac{FP}{FP+TN}$)

hemos calculado diferentes métricas de ML: accuracy ($\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$), precision ($\frac{\text{TP}}{\text{TP + FP}}$), recall ($\frac{\text{TP}}{\text{TP + FN}}$), f_1-score ($2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$), and balanced accuracy [3] ($\frac{1}{2}\left(\frac{\text{TP}}{\text{TP + FN}} + \frac{\text{TN}}{\text{TN + FP}}\right)$), respectivamente.

Dado que SHAP (SHapley Additive exPlanations) [18] es una de las técnicas de explicabilidad de extracción de características más conocidas y probadas, hemos hecho uso de ésta para extraer los pesos del modelo de ML utilizado. Esta técnica analiza las variables del modelo y la influencia de cada una de éstas sobre las salidas de dicho modelo. Para poder extraer los pesos globales en lugar de locales a través de SHAP, hemos seguido su documentación oficial y aplicado el valor medio absoluto [19]. Para el conjunto de datos de Adult, la variable a la que el modelo le ha dado más peso es el estado civil de la persona ("marital.status"), con un peso de 0.895.

Equidad:

Ya que el conjunto de datos de Adult separa grupos de individuos entre diferentes atributos protegidos, hemos seleccionado la equidad de grupo. En este contexto, hemos definido dos métricas de equidad basadas en el concepto de probabilidades igualadas, donde se tiene en cuenta tanto el ratio de positivos como el de negativos cubriendo así ambos escenarios. Más en concreto, hemos definido las métricas de: (i) Ratio de Verdaderos Ponderado (*RVP*), donde nos centramos en los resultados positivos y (ii) Ratio de Falsos Ponderado (*RFP*), donde nos centramos en los resultados negativos. Ambas métricas están enfocadas en ayudar a evaluar diferentes tipos de conjuntos de datos, incluso pequeños o desbalanceados. Además, sendas métricas son calculadas para los dos grupos o categorías que son incluidas en cada combinación, tomando así en cuenta el número de filas de cada una de éstas. Las ecuaciones 1 y 2 muestran cómo calcular dichas métricas, representando los dos grupos distintos como "0" y "1".

$$RVP = (TPR_0 + TNR_0) \cdot \frac{rows_0}{rows} + (TPR_1 + TNR_1) \cdot \frac{rows_1}{rows}, \tag{1}$$

$$RFP = (FPR_0 + FNR_0) \cdot \frac{rows_0}{rows} + (FPR_1 + FNR_1) \cdot \frac{rows_1}{rows}, \tag{2}$$

donde para cada grupo *TPR* representa el Ratio de Verdaderos Positivos, *TNR* representa el Ratio de Verdaderos Negativos, *FPR* representa el Ratio de Falsos Positivos, *FPR* representa el Ratio de Falsos Negativos, y *rows* representa el número total de filas de la combinación, *i.e.*, $rows = rows_0 + rows_1$.

De esta forma, los valores de los ratios estarán comprendidos en el intervalo cerrado [0,1] y las divisiones entre las filas de cada grupo y el total de filas también estarán comprendidas entre dicho intervalo cerrado [0,1]. En consecuencia, las ecuaciones 2 (*RFP*) y 1 (*RVP*) tendrán un valor final de: $\frac{[0,2]}{[0,1]} + \frac{[0,2]}{[0,1]}$, dónde el

Tabla 1. Métricas de los modelos de ML que se han empleado para evaluar nuestra propuesta.

	Accuracy	**Precision**	**Recall**	**F1-Score**	**Balanced Accuracy** [3]
Adult	0.836	0.697	0.581	0.634	0.750

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

segundo y el cuarto elemento son inversamente proporcionados, *i.e.*, la suma de ambos es igual a uno. Así, el valor final de *RVP* y *RFP* estará comprendido en el intervalo cerrado [0,2].

En este contexto, para *RVP*, el mejor valor (más justo) será 2 pues se acierta siempre, y el peor (menos justo) 0 que representaría que no se ha acertado nunca. De forma contraria, para *RFP*, el mejor valor (más justo) es 0 (no se ha fallado nunca) y el peor (menos justo) es 2 pues representaría que se ha fallado siempre.

Entrando en detalle en el conjunto de datos de Adult, hay más de un **atributo protegido** que debe ser analizado. Debido a esto, hemos afrontado la interseccionalidad de múltiples atributos protegidos ejecutando el algoritmo ADE para cada uno de dichos atributos, por lo que se han generado diferentes AAE dependiendo del atributo protegido de cada ejecución. Para el ejemplo que se detalla, hemos detectado la variable sexo ("Sex") como atributo protegido. Una vez éste ha sido detectado, hemos incluido esta variable en el algoritmo ADE. Por consiguiente, podemos obtener información acerca del comportamiento de la equidad teniendo dicha variable en cuenta. Dicha información será de utilidad en el proceso de análisis y toma de decisiones.

3.2. Algoritmo: Árbol de Detección de Equidad (ADE)

Para la ejecución del algoritmo, la entrada ha consistido en los puntos presentados anteriormente. La principal entrada es el conjunto de datos de Adult categorizado, el cual incluye las clases predicha y la real como variables con las que el algoritmo ADE trabaja (260KB). Como salida, el AAE generado (un archivo .json) tiene un tamaño de 204KB. Dicho AAE se analiza a continuación.

3.3. Salida: Árbol de Análisis de Equidad (AAE)

Se procede a detallar la salida del ADE, es decir, el AAE. Dicha información ha sido obtenida en el conjunto de datos de Adult y el atributo protegido "Sexo". De esta forma, se presenta la Tabla 2, donde se observan la distribución de filas y las tasas de acierto y fallo para los diferentes casos en los grupos 4 y 2 del estado civil, viudo/a y nunca habiendo contraído matrimonio, respectivamente. Además, dentro del grupo 2, se presenta la misma información para la variable protegida "Sexo", donde 0 representa a las mujeres y 1 a los hombres.

Analizando esta información, se puede observar que la suma de los ratios positivos es mayor dónde hay más filas, de la misma forma que la suma de ratios

Tabla 2. Distribución de los datos en el conjunto de datos de Adult.

	Filas	***TPR***	***FPR***	***TNR***	***FNR***
2 (nunca casada/o)	3721	0.666	0.215	0.785	0.334
2.0 (mujer nunca casada)	3298	0.686	0.277	0.723	0.314
2.1 (hombre nunca casado)	423	0.663	0.207	0.793	0.337
4 (persona viuda)	2641	0.254	0.005	0.995	0.746

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

Tabla 3. Resultados de los experimentos aplicados en el conjunto de datos de Adult.

	***RFP*↓**	***RVP*↑**	**Acc.**	**Precision**	**Recall**	F_1	**Balanced acc.** [3]
Original	0.756	1.244	0.836	0.697	0.581	0.634	0.750
SMOTE	0.530↓	1.471↑	0.847↑	0.725↑	0.667↑	0.695↑	0.789↑
Undersampling	0.539↓	1.461↑	0.848↑	0.731↑	0.665↑	0.696↑	0.789↑
Oversampling	0.537↓	1.463↑	0.856↑	0.727↑	0.652↑	0.687↑	0.787↑

negativos es menor en esos grupos, es decir, el número de filas puede ser la razón principal que afecta a las métricas. Por otro lado, pese que hay más filas para mujeres nunca casadas que para hombres nunca casados, se observa como se está sobrestimando a los hombres sobre las mujeres.

Para ejemplificar cómo esta información puede ser usada para tomar medidas correctivas, se ha hecho uso de técnicas de rebalanceo. Como la Tabla 3 muestra, dichas técnicas (SMOTE [6], Undersampling [11], y Oversampling [16]) nos han llevado a mejorar los resultados en las métricas de equidad (menor *RFP* significa mejora, y mayor *RVP* significa mejora). Por otro lado, también se observa cómo el valor original ha sido mejorado para las métricas de ML: accuracy, precision, recall, f_1-score y balanced accuracy [3].

Finalmente, nuestra propuesta también se ha aplicado a otros cuatro conjuntos de datos, haciendo junto a Adult [9] un total de **cinco conjuntos de datos en 4 dominios distintos**. Más en concreto, esos conjuntos de datos son: Statlog (German Credit Data) [14], COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [20], Law school [27], y diabetes [26]. En la Tabla 4 se presenta la información de dichos conjuntos de datos. Debido a limitaciones de espacio, sólo se ha detallado el conjunto de datos de Adult. No obstante, a continuación se presenta un resumen del resto de conjuntos de datos y cómo se ha aplicado la propuesta junto a los resultados de ello.

Detallando los experimentos ejecutados para el resto de conjuntos de datos incluidos en este trabajo, de la misma manera que con el conjunto de Adult, se han aplicado técnicas de rebalanceo y después se ha vuelto a entrenar y evaluar cada modelo. Además, para aportar mayor fiabilidad a los resultados, se han ejecutado todos los experimentos diez (10) veces y calculado su media. De esta forma, se ha observado como tanto las métricas de ML como las métricas de equidad han mejorado en cuatro de los cinco conjuntos de datos utilizados. En los experimentos donde no se ha mejorado, nuestra propuesta ha facilitado datos

Tabla 4. Información de los conjuntos de datos incluidos en nuestros experimentos

	Instancias	**Attrs.**	**Dominio**	**Attr.Prot.**	**Clase salida**
Adult	32.561	14	Finanzas	Sexo, raza, edad	Income
Statlog	1.000	20	Finanzas	Sexo, edad	Credit score
COMPAS	7.214	52	Criminal	Sexo, raza , edad	Two year recid.
Diabetes	101.766	49	Sanitario	Sexo, raza, edad	Readmit
Law school	18.692	11	Educación	Sexo, raza	Pass exam

para saber dónde se encuentra el problema. No obstante, el rebalanceo no ha sido la solución adecuada y habría que buscar soluciones alternativas que podrían apoyarse en la información que proporcionamos gracias a nuestra propuesta.

4. Estado de la cuestión

En primer lugar, dado que nuestra propuesta se enfoca en la evaluación de la equidad, es imprescindible detallar el posicionamiento de las métricas propuestas dentro del contexto del resto de métricas. La mayoría de éstas están basadas en interpretar la precisión del modelo y aplicar ratios [13]. Otros trabajos han estudiado aproximaciones existentes para trabajar con atributos protegidos (*e.g.*, [5]). No obstante, estas técnicas y métricas pueden no ser útiles y confiables en conjuntos de datos reducidos y/o desbalanceados.

En segundo lugar, como se argumenta en [22], basarse en aproximaciones jerárquicas permite la extracción flexible de grupos y subgrupos en los datos. De acuerdo al análisis de los autores, la elección de discretizar los datos arrojó luz en los datos, haciendo éstos más comprensibles. En este contexto, varios trabajos recientes se enfocaron en identificar grupos -y subgrupos- problemáticos con resultados distintos para evaluar la equidad. Por ejemplo, FairVis [4] afronta este problema usando algoritmos de grupos (*clustering*). Por otro lado, SliceFinder [8] y la herramienta de análisis de error [1] derivan subgrupos a través de métodos basados en árboles. Además, SliceLine [25] presenta un algoritmo de enumeración basado en álgebra lineal y DivExplorer [21] identifica subgrupos asociados a comportamientos problemáticos en conjuntos de objetos. Sin embargo, ninguna de estas aproximaciones centra la solución en guiar el proceso y proporcionar un análisis no sólo entre grupos sino también intragrupo para las características de los datos en un sentido explicable, lo que permite a los expertos en ML observar fácil y rápidamente la información en los procesos de análisis.

Finalmente, en el contexto de XAI, un trabajo reciente trata de enfrentarse al problema de la injusticia a través de la aplicación de técnicas de explicabilidad. En [23], los autores buscan patrones para explicar el comportamiento de la injusticia en los datos. De esta forma, este trabajo se centra en las relaciones entre los diferentes atributos de los datos. Sin embargo, su propuesta no guía el proceso de análisis ni explora un análisis intragrupo (sólo se centra en el análisis entre grupos), lo que es necesario para poder conocer dónde se deben tomar medidas correctivas.

En resumen, hay diferentes aproximaciones para afrontar el problema de injusticia en los sistemas de IA y modelos de ML, algunas de las cuales incluyen la aplicación de métodos de explicabilidad. Diferenciándonos de éstas, en este artículo hemos presentado un proceso para evaluar la equidad incluso en conjuntos de datos reducidos y/o desbalanceados, además de guiar el proceso y aportar un análisis no sólo entre grupo sino también intragrupo.

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

5. Conclusiones y trabajo futuro

En este artículo, hemos presentado una propuesta que permite a los expertos en ML conocer si un sistema de IA está tomando o no decisiones injustas. Nuestra propuesta identifica qué grupos podrían estar siendo afectados por un comportamiento discriminatorio sin previo conocimiento del dominio de aplicación en cuestión. El proceso propuesto se basa en un novedoso algoritmo jerárquico, llamado Árbol de Detección de Equidad (ADE), y diferencia a modo de guía qué se debe hacer para usar el algoritmo ADE (la entrada), el algoritmo en sí, y la salida que emitirá. Un aspecto clave del algoritmo ADE es que funciona recursivamente, por lo que nos permite explorar rápidamente el conjunto de datos y analizar exhaustivamente los grupos y subgrupos que normalmente no serían fácilmente detectados. Gracias a esto, el algoritmo ADE proporciona un análisis jerárquico y no sólo entre grupos sino también intragrupo y guarda toda la información en el Árbol de Análisis de Equidad (la salida anteriormente mencionada: AAE), lo que ayuda a los expertos en ML a analizar los datos, incluso en escenarios complejos donde hay muchas variables.

Además, para poder evaluar la equidad incluso en conjuntos de datos reducidos y/o desbalanceados, hemos propuesto dos nuevas métricas de equidad: Ratio de Verdaderos Ponderado (*RVP*) y Ratio de Falsos Ponderado (*RFP*).

Para probar la aplicabilidad de nuestra propuesta, hemos ejecutado una serie de experimentos en cinco ampliamente conocidos conjuntos de datos, los cuales cubren diferentes contextos y situaciones, incluyendo casos donde hay un número reducido de filas para ciertos grupos. Tras ello, para poder ejemplificar posibles acciones correctivas, hemos experimentado usando técnicas de rebalanceo. Observando los resultados obtenidos en el conjunto de datos de Adult, comprobamos cómo el modelo de ML tiende a sobrestimar (i) a los hombres en prejuicio de las mujeres y (ii) a los grupos con más filas. Motivado por ello, se han ejecutado una serie de experimentos que nos han mostrado cómo es posible hacer uso de la información que aporta nuestra propuesta para mejorar los modelos de ML tanto en métricas de ML como de equidad, lo que es esencial para implementar sistemas de IA.

Finalmente, este artículo sienta las bases para obtener más información que permita corregir los desequilibrios en los conjuntos de datos y evitar que pasen desapercibidos. Más investigación es necesaria para poder continuar progresando en esta línea, evaluando la equidad en distintos contextos e incluyendo y enfatizando no sólo la IA ética sino también la IA explicable.

Agradecimientos

Este trabajo ha sido cofinanciado por: (i) el proyecto AETHER-UA (PID2020-112540RB-C43), financiado por el Ministerio de Ciencia e Innovación; y (ii) el proyecto BALLADEER (PROMETEO/ 2021/ 088), financiado por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital (Generalitat Valenciana).

Referencias

1. Error analysis. `https://erroranalysis.ai/`, accessed: 2023/10/04
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
3. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. pp. 3121–3124 (2010). `https://doi.org/10.1109/ICPR.2010.764`
4. Cabrera, Á.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., Chau, D.H.: Fairvis: Visual analytics for discovering intersectional bias in machine learning. In: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 46–56 (2019)
5. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. Scientific Reports **12**(1), 4209 (2022)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Jo. of Artificial Intellig. research **16**, 321–357 (2002)
7. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
8. Chung, Y., Kraska, T., Polyzotis, N., Tae, K.H., Whang, S.E.: Slice finder: Automated data slicing for model validation. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). pp. 1550–1553 (2019)
9. Dua, D., Graff, C., et al.: Uci machine learning repository (2017)
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
11. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation **17**(3), 275–306 (2009)
12. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)
13. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
14. Hofmann, H.: Statlog (german credit data) data set. UCI Repository of Machine Learning Databases **53** (1994)
15. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning. pp. 2564–2572. PMLR (2018)
16. Kovács, G.: An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Applied Soft Computing **83**, 105662 (2019)
17. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery p. e1452 (2022)

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58

18. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
19. Lundberg, S.M., Lee, S.I.: An introduction to explainable ai with shapley values (2020), https://shap.readthedocs.io/en/latest/index.html, accessed: 2023/11/16
20. Office, B.C.C., Office, B.C.S., of Corrections, F.D., ProPublica: Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis (2021), accessed: 2024/02/01
21. Pastor, E., de Alfaro, L., Baralis, E.: Looking for trouble: Analyzing classifier behavior via pattern divergence. In: Proceedings of the 2021 International Conference on Management of Data. p. 1400–1412. SIGMOD '21, Association for Computing Machinery, New York, NY, USA (2021)
22. Pastor, E., Baralis, E., de Alfaro, L., et al.: A hierarchical approach to anomalous subgroup discovery. In: 39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, California, USA, April 3–7, 2023. IEEE (2023)
23. Pradhan, R., Zhu, J., Glavic, B., Salimi, B.: Interpretable data-based explanations for fairness debugging. arXiv preprint arXiv:2112.09745 (2021)
24. Prasanth Kadiyala, S., Woo, W.L.: Flood prediction and analysis on the relevance of features using explainable artificial intelligence. In: 2021 2nd Artificial Intelligence and Complex Systems Conference. p. 1–6. AICSconf '21, Association for Computing Machinery, New York, NY, USA (2022)
25. Sagadeeva, S., Boehm, M.: Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In: Proceedings of the 2021 International Conference on Management of Data. p. 2290–2299. SIGMOD '21, Association for Computing Machinery, New York, NY, USA (2021)
26. Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., Clore, J.N.: Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed research international (2014)
27. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series. (1998)

Navarro, Á., Lavalle, A., Maté, A., Trujillo, J.: Proceso para la evaluación de la equidad en sistemas de Inteligencia Artificial. In: Rodríguez Luaces, M. A. (ed.) Actas de las XXVIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2024). Sistedes (2024). https://hdl.handle.net/11705/JISBD/2024/58