

DETECCIÓN DE DISCRIMINACIÓN EN UN MODELO DE RED NEURONAL

Salaberry, Natalia¹

¹Universidad de Buenos Aires, Facultad de Ciencias Económicas, Centro de Investigación en Métodos Cuantitativos Aplicados a la Gestión y la Economía (CMA), Av. Córdoba 2122- 1120AAQ, Ciudad Autónoma de Buenos Aires, República Argentina

natalia.salaberry@economias.uba.ar

Resumen

Recibido: 28-08-2025

Aceptado: 03-03-2026

Palabras clave

Análisis de Equidad,
Red Neuronal
Clasificación
Créditos.

Este trabajo tiene por objetivo la aplicación de una metodología de análisis de equidad para alcanzar una aproximación a la detección cuantitativa de discriminación de un modelo de inteligencia artificial. Particularmente, se busca poder detectar si un modelo de red neuronal reproduce un sesgo humano discriminatorio contenido en los datos utilizados. Dado su alto rendimiento para la obtención de resultados eficaces, poder comprender si este tipo de modelos discrimina por grupos sociales permite contribuir a realizar un uso responsable. Por este último, se define a la confiabilidad en los resultados obtenidos necesaria para garantizar una transparencia de cara a la sociedad. En este contexto, la aplicación de una metodología propuesta - mediante un caso de aplicación - resulta oportuno para poder mostrar cómo es posible realizar un análisis de equidad en una red neuronal. En función de los resultados obtenidos se logra mostrar como el modelo reproduce un sesgo discriminatorio por género. Una vez detectado el problema, se propone una corrección de este alcanzando resultados favorables para evitar la no reproducción del sesgo discriminatorio. De esta manera, se logra visibilizar la responsabilidad involucrada de quien utiliza este tipo de modelo y cómo es posible contribuir al objetivo de transparencia dentro de un marco de uso responsable de la inteligencia artificial.

EQUITY ANALYSIS IN A NEURAL NETWORK MODEL. APPLICATION CASE FOR CLASSIFICATION IN THE GRANTING OF LOANS.

Abstract

Keywords

Equity Analysis,
Neuronal Network,
Classification,
Credit.

This work aims to apply a fairness analysis methodology to achieve a quantitative approximation for discrimination detection in an artificial intelligence model. Specifically, it seeks to detect whether a neural network model reproduces a discriminatory human bias contained in the data used. Given its high performance in obtaining effective results, understanding whether this type of model discriminates based on social groups allows us to contribute to its responsible use. The latter is defined as the reliability of the results obtained, necessary to guarantee transparency for society. In this context, the application of a proposed methodology—through a case study—is opportune to demonstrate how it is possible to perform a fairness analysis in a neural network. Based on the results obtained, it is possible to demonstrate how the model reproduces a discriminatory gender bias. Once the problem is detected, a correction is proposed, achieving favorable results to avoid the non-reproduction of discriminatory bias. In this way, the responsibility involved in those who use this type of model is made visible and how it is possible to contribute to the objective of transparency within a framework of responsible use of artificial intelligence.

INTRODUCCIÓN

El avance en el uso de tecnologías de inteligencia artificial (IA) para la resolución de problemas ha mostrado la posibilidad de obtener resultados altamente eficientes. Sin embargo, se han suscitado discusiones acerca de cómo evitar la incorporación de sesgos humanos que puedan conducir a un uso injusto de la información. A partir de ello, diferentes organismos internacionales como la OCDE y la UNESCO han desarrollado ciertos lineamientos sobre que es necesario tener en cuenta a la hora de utilizar este tipo de tecnología debido al impacto social y cultural que puede causar. Argentina, no ha hecho la vista a un lado y también ha emitido cierta disposición como marco para el uso responsable de la IA.

Ante este escenario, surge la necesidad de poder realizar un análisis de equidad a partir de detectar si entre los datos utilizados hay una variable sensible que pueda derivar en discriminación. Una vez aplicado un modelo de red neuronal, se busca poder realizar una evaluación de si este reproduce el sesgo humano o no. Desde un enfoque cuantitativo, se han desarrollado metodologías que permiten llevar a cabo dicha tarea. En este contexto, el presente trabajo tiene por objetivo detectar si una red neural está siendo justa o no frente a datos que se encuentran desbalanceados según una variable sensible como lo es el género.

En este contexto, surge el siguiente interrogante a responder: ¿Cómo es posible detectar si una red neuronal es justa o no desde un enfoque cuantitativo? Para poder aproximar una respuesta, en primer lugar, se aborda el concepto de gobernanza de modelos. En el primer apartado se desarrolla una introducción a este, arribando a la conclusión de que es necesario establecer un marco de responsabilidad en el uso de inteligencia artificial. El impacto socio cultural que este tipo de tecnología trae aparejado en la medida que sus resultados son utilizados para la toma de decisiones puede traer consecuencias que alteren los valores sociales y derechos humanos. A partir de ello, surge la necesidad de realizar un uso de la IA bajo un marco de responsabilidad que incluye poder brindar garantías de que el modelo utilizado es justo para los individuos.

En el segundo apartado, se introduce el modelo de una red neuronal secuencial para resolver un problema de clasificación. Si bien se trata de un modelo matemático probabilístico cuyos fundamentos principales son presentados, requiere contar con un importante conocimiento especializado para lograr comprender si está reproduciendo un sesgo humano. Por esta razón, más adelante se realiza una aplicación de caso para entonces poder realizar un análisis de equidad.

Para poder llevar a cabo un análisis de equidad de una red neuronal, en el tercer apartado se presenta una metodología cuantitativa. Se trata de una metodología que brinda la posibilidad de obtener medidas para detectar si un modelo discrimina o no según una variable sensible como,

por ejemplo, el género. En caso de que las medidas involucradas muestren un accionar injusto del modelo, se deberá adoptar alguna medida correctiva y realizar nuevamente su aplicación.

En el cuarto apartado, se lleva a cabo la aplicación de un modelo de red neuronal sobre un conjunto de datos con el objetivo de predecir si los clientes presentan riesgo de *default* o no a través de *Python*. Como resultado se obtiene que el rendimiento mostrado por la red neuronal es eficaz a este fin. A pesar de este resultado, surge la necesidad de poder evaluar si el modelo está siendo justo o no según el género del cliente.

Finalmente, en el último apartado se lleva a cabo la aplicación de una metodología para la detección de discriminación según género mediante *Python*. Del análisis realizado, se detecta que el modelo es inequitativo. De este modo se habilita la posibilidad de realizar un ajuste en la muestra de datos para el entrenamiento de este con el fin de subsanar el problema. En base al conjunto de resultados obtenidos, se logra mostrar cómo es posible evitar que un modelo reproduzca un sesgo humano de discriminación, constituyéndose en una contribución para un uso responsable de la IA.

1. GOBERNANZA DE MODELOS DE INTELIGENCIA ARTIFICIAL

Según Dignum (2019) la inteligencia artificial (IA) puede definirse como un sistema informático que procesa datos con cierta inteligencia para producir información. Tal inteligencia se encuentra asociada al concepto de inteligencia humana desde una perspectiva de las ciencias de computación. En este sentido, surgen dos aspectos característicos. Por un lado, el pensamiento asociado al proceso cognitivo, es decir, al modelo de aprendizaje. Por el otro, el razonamiento lógico para la modelización de la incertidumbre y posibilidad de manejo de la complejidad.

La instrumentación informática de este tipo de modelos se denomina algoritmo. Un algoritmo es una secuencia de instrucciones elementales ejecutables, generalmente destinadas a lograr un propósito específico (Erickson, 1999; Hromkovič, 2004). De este modo, su objetivo es que las computadoras puedan ejecutar instrucciones humanas y aprender de la experiencia obtenida para mejorar su rendimiento a la vez que su complejidad crece (Kolanovic y Krishnamachari, 2017; Dignum, 2019).

La complejidad de las tecnologías de inteligencia artificial (IA) ha permitido, por un lado, obtener resultados superadores para la resolución de diferente tipo de problemas en base a datos; por el otro, ha despertado diferentes cuestionamientos en torno a su uso. Entre este tipo de modelos se encuentran las redes neuronales. Se caracterizan por realizar un proceso de aprendizaje simulando el comportamiento del cerebro humano. Actúan por etapas a partir de datos de entrenamiento, hasta que encuentran la mejor relación de ajuste entre las variables para obtener una predicción más acertada (Kolanovic y Krishnamachari, 2017; Dignum, 2019).

La complejidad intrínseca de los modelos de redes neuronales ha puesto de relieve la necesidad de poder ejercer un control analítico de su operatoria para contribuir a un uso de manera responsable. Diferentes organismos internacionales han emitido ciertos principios entorno a la utilización de este tipo de tecnologías, como ser la Organización para la Cooperación y el Desarrollo Económico (OCDE) en 2019¹ o la UNESCO (*United Nations Educational, Scientific and Cultural Organization*) en 2022². Estos principios tienen por objetivo establecer estándares sobre el uso confiable de la IA en pos del respeto de los derechos humanos y los valores sociales y democráticos. A partir de ello, se busca que un modelo no solo sea utilizado para obtener el “que” (resolver un problema a través de obtener un valor predicho) sino que es necesario saber el “porque” (cómo el modelo actúa frente a determinados casos o posibles sesgos).

En esta misma línea, en el caso particular de Argentina, el 2 de junio de 2023 a través de la Jefatura de Gabinete de Ministros en conjunto con la Secretaría de Tecnologías de la Información, se emitió la disposición 2/2023 sobre “Recomendaciones para una Inteligencia Artificial Fiable”. Esta tiene por finalidad brindar “un marco para la adopción de Inteligencia Artificial centrada en el ciudadano y sus derechos, concibiendo su aspecto social y estratégico, asegurando un óptimo funcionamiento de la prestación de servicios y un enfoque ético” (Anexo I, p.3) con el fin de lograr que el uso de la IA sea transparente, equitativo y responsable. Particularmente, para el sector financiero, el BCRA (Banco Central de la República Argentina) emite en marzo de 2023 la normativa A7724³ que busca regular la utilización de la IA en organizaciones financieras.

Este marco parte de una concepción fundamental que es que la responsabilidad de la toma de decisiones –utilizar o no IA - recae en el ser humano, mientras que la IA es el instrumento que ejecuta instrucciones. En este sentido resulta necesario evaluar el impacto que los resultados obtenidos pueden tener en el conjunto de la sociedad. Una primera medida es hacer explicables los modelos con el fin de poder comprender su funcionamiento (Ali et al., 2023)⁴. Al mismo tiempo resultará necesario evaluar su transparencia. Particularmente, se busca realizar una evaluación de equidad del modelo (Brzezinski et. al, 2019; Mehrabi et. al, 2019; Pessach y Shmueli, 2022), es decir, si es justo o no frente a grupos. En conjunto permitirán actuar con responsabilidad en el sentido de poder construir confianza sobre el uso de un modelo de IA, evitándose la incorporación de sesgos discriminatorios humanos. De aquí que la gobernanza de modelos es

¹ Que entre los países miembros adheridos se encuentra la República Argentina

² Se utiliza el término ética para referirse a la responsabilidad necesaria que debe tenerse en cuenta a la hora de utilizar tecnología de inteligencia artificial.

³ Véase <https://www.bcra.gob.ar/pdfs/comytexord/A7724.pdf>

⁴ Véase Salaberry, N. (2024) “LA INTERPRETABILIDAD CUANTITATIVA DE UNA RED NEURONAL. CASO DE APLICACIÓN DE LA METODOLOGÍA DE SHAP VALUES” REVISTA DE INVESTIGACIÓN EN MODELOS MATEMÁTICOS APLICADOS A LA GESTIÓN Y LA ECONOMÍA – AÑO 10 VOLUMEN I (2024-I) <https://www.economicas.uba.ar/investigacion/wp-content/uploads/Salaberry.pdf>

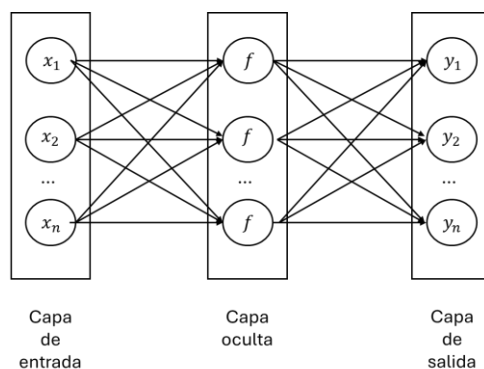
parte fundamental de la construcción de integridad en el contexto de toma de decisiones basada en información.

De este modo, un buen gobierno de modelos podría decirse que posee tres pilares fundamentales: las personas, los procesos y la tecnología. A partir de ello, se establece un ciclo funcional que comienza por la identificación de desafíos, con el fin de establecer el impacto en torno a los aspectos sociales. En una segunda etapa, se desarrollan políticas y procedimientos alineados al cumplimiento del marco establecido. Luego se procede con la ejecución de tales políticas con el fin de lograr los objetivos planteados. Finalmente, se realiza monitoreo y medición de resultados. Este ciclo es de naturaleza continúa debido a la propia dinámica del ciclo de vida de la IA en tanto instrumento para la generación de información confiable para la toma de decisiones.

2. EL MODELO: RED NEURONAL SECUENCIAL PARA CLASIFICACIÓN

Una red neuronal secuencial es un modelo de abstracción de la interconexión de neuronas cerebrales para ser implementadas a través de una computadora (Wang y Fu, 2005). Para la resolución de problemas de clasificación, la red típica que se suele utilizar es un perceptrón multicapa. Esta consta de una capa de entrada por donde ingresan los vectores de valores, al menos una capa oculta o intermedia donde se procesan los valores anteriores a través una función de optimización dentro de los nodos (o neuronas) interconectados y una capa de salida donde se obtiene el valor de la clase predicha por el modelo. En la figura a continuación se esquematiza lo anterior para una red con una sola capa oculta.

Figura 1: esquema de una red neuronal multicapa



Fuente: elaboración propia en base a Wang y Fu (2005)

En la figura 1, cada x_i representa un vector de valores de entrada (los valores que toma cada predictora), f es la función de optimización de los pesos y y_j son las salidas predichas obtenidas⁵. De este modo, la predicción viene dada por:

⁵ En el caso de una clasificación binaria, el valor que puede tomar y_j es 0 o 1.

$$y_j = f(\sum_{i=1}^K W_{ij}^{(2)} \phi_i(x) + b_j^{(2)}) \quad (1)$$

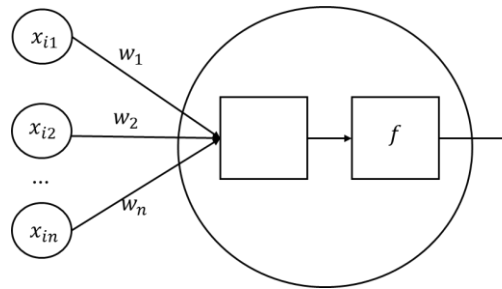
La ecuación (1) muestra la operación matemática realizada por el modelo de red neural secuencial para obtener una predicción. En esta, W_{ij} son los pesos sinápticos, $\phi_i(x)$ es el resultado obtenido en la neurona en la capa intermedia a partir del valor de entrada (x) y b_j es el sesgo (o desvío o error) respecto del valor real de la clase.

Luego, $\phi_i(x)$ viene dado por:

$$\phi_i(x) = f(W_i^{(1)} \cdot x + b_i^{(1)}) \quad (2)$$

En la ecuación (2), $W_i^{(1)}$ es el vector de pesos que conecta el vector de entrada x con la neurona i en la capa intermedia, y $b_i^{(1)}$ es el sesgo que se produce en la misma neurona. Si se hiciera un zoom en una neurona del esquema de la figura 1, es posible representar esquemáticamente la operación realizada por la ecuación (2):

Figura 2: proceso en una neurona



Fuente: elaboración propia en base a Wang y Fu (2005)

En la Figura 2, se puede observar que los valores de entrada ingresan a la neurona de la capa intermedia y a partir de un proceso de cálculo en función de los pesos emite un resultado o salida. La función f es una función de optimización. En el caso de los problemas de clasificación, la más frecuentemente utilizada es la función sigmoide logística que viene dada por:

$$f(z) = \frac{1}{1+e^{-\beta x}} \quad (3)$$

La optimización de los parámetros a través de la función sigmoidea logística dada por la ecuación (3), se alcanza mediante un algoritmo de *Back-Propagation*⁶. Esencialmente este algoritmo implica realizar optimización hacia atrás de los pesos para encontrar el peso óptimo en cada neurona para

⁶ Para más acerca de Back-Propagation véase <http://neuralnetworksanddeeplearning.com/chap2.html>

cada valor de entrada de una manera iterativa. Una vez que ya no se logra optimizar más los pesos, entonces se produce la salida, es decir, la predicción.

Como es posible notar, la complejidad de funcionamiento de una red neuronal requiere de un conocimiento muy especializado por lo que en cierto modo no resulta fácilmente accesible realizar su comprensión. Más aún si se quiere determinar si el modelo ha cometido discriminación en base a los datos con los cuales fue entrenado. Por esta razón, cuando en el apartado anterior se indicaba la necesidad de poder hacer una evaluación de equidad, surge la necesidad de contar con alguna metodología que aproxime a poder lograrlo. En el siguiente apartado se aborda una propuesta.

3. METODOLOGÍA PARA ANÁLISIS DE EQUIDAD

El análisis de equidad de un modelo implica poder brindar cierta garantía de que el resultado obtenido no perjudique (para bien o mal) a ningún grupo poblacional en función de variables sensibles (Mehrabi et. al, 2019; Pessach y Shmueli, 2022). Particularmente, para los modelos de clasificación, las medidas que permiten realizar un análisis de equidad se construyen partiendo de los elementos contenidos en una matriz de confusión. Esta matriz muestra como el modelo aplicado predice correctamente o no cada caso:

Tabla 1: Matriz de Confusión

	Positivos	Negativos
Positivos	Verdaderos positivos (TP)	Falsos positivos (FP)
Negativos	Falsos negativos (FN)	Verdaderos negativos (TN)

Fuente: elaboración propia

En la matriz de confusión mostrada en la Tabla 1, los casos TP y TN, son aquellos casos predichos por el modelo de manera correcta. Para la clasificación binaria, esto es: si era 0 el valor fue predicho como 0 (TP); si era 1 fue predicho como 1 (TN). En cambio, los casos FN y FP son aquellos predichos incorrectamente por el modelo. A su vez, la matriz de confusión puede conformarse como la suma de la matriz correspondiente a los casos pertenecientes al grupo mayoritario y al minoritario. De este modo se tendrá:

Grupo mayoritario	Positivos	Negativos	Grupo minoritario	Positivos	Negativos
Positivos	TP _p	FP _p	Positivos	TP _M	FP _M
Negativos	FN _p	TN _p	Negativos	FN _M	TN _M

A partir de separar la matriz de confusión por grupos, es posible establecer tres grupos de medidas de equidad. Siguiendo a Dariusz Brzezinski et. al (2019), se parte de considerar el atributo minoritario (S), la variable objetivo (Y) y el score de clasificación (C) que representa la puntuación predicha por el modelo en el rango [0;1]. De este modo, los autores definen tres grupos de medidas para abordar la no discriminación: la independencia entre C y S, la separación entre C y S dado Y la suficiencia entre Y S dado C. En el grupo de la independencia se encuentran las siguientes dos medidas:

$$\textbf{Igualdad de precisión} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{TP_P+TN_P}{TP_P+FP_P+TN_P+FN_P} - \frac{TP_M+TN_M}{TP_M+FP_M+TN_M+FN_M} \quad (4)$$

$$\textbf{Ratio de paridad} = \frac{TP+FP}{TP+FP+FN+TN} = \frac{TP_P+FP_P}{TP_P+FP_P+TN_P+FN_P} - \frac{TP_M+FP_M}{TP_M+FP_M+TN_M+FN_M} \quad (5)$$

donde la ecuación (4) es la $\mathbf{P(C = Y/S)}$, es decir, exactitud de la probabilidad predicha por grupo; y la ecuación (5) es $\mathbf{P(C = 1/S)}$, es decir, la probabilidad predicha para el valor 1 de la clase objetivo en cada grupo. En el segundo grupo de medidas, el de separación, se encuentran las siguientes dos medidas:

$$\textbf{Igualdad de oportunidades} = \frac{TP}{TP+FN} = \frac{TP_P}{TP_P+FN_P} - \frac{TP_M}{TP_M+FN_M} \quad (6)$$

$$\textbf{Igualdad predictiva} = \frac{FP}{FP+TN} = \frac{FP_P}{FP_P+TN_P} - \frac{FP_M}{FP_M+TN_M} \quad (7)$$

donde la ecuación (6) es la $\mathbf{P(C = 1/Y = 1)}$, es decir, la probabilidad predicha para el valor 1 de la clase objetivo dado que el verdadero valor de esta era 1; y la ecuación (7) es $\mathbf{P(C = 1/Y = 0)}$, es decir, los casos incorrectamente clasificados debido a que el modelo predijo una probabilidad de que sea 1 cuando el verdadero valor de la clase era 0. Por último, en el tercer grupo de medidas, el de suficiencia, se encuentran las siguientes dos medidas:

$$\textbf{Paridad predictiva positiva} = \frac{TP}{TP+FP} = \frac{TP_P}{TP_P+FP_P} - \frac{TP_M}{TP_M+FP_M} \quad (8)$$

$$\textbf{Paridad predictiva negativa} = \frac{TN}{FN+TN} = \frac{TN_P}{FN_P+TN_P} - \frac{TN_M}{FN_M+TN_M} \quad (9)$$

donde la ecuación (8) es la $\mathbf{P(Y = 1/C = 1)}$, siendo la probabilidad de casos correctamente clasificados como 1 cuando su valor era 1; y la ecuación (9) es $\mathbf{P(Y = 0/C = 0)}$, es decir, la probabilidad de los casos correctamente clasificados como 0.

El criterio para determinar si un modelo discrimina según una variable identificada como sensible, consiste en obtener al menos dos de las medidas establecidas que superen un umbral de 0.8. Este valor es un *score* construido para la visualización de las medidas. Consiste en el cociente entre el

valor obtenido de cada medida calculada para el grupo minoritario y el grupo mayoritario. Con el fin de poner a prueba la metodología presentada en este apartado, a continuación, se desarrolla un caso de aplicación mediante la utilización del software Python⁷.

4. CASO DE APLICACIÓN: RED NEURONAL PARA LA CLASIFICACIÓN DE RIESGO CREDITICIO

Llevar a cabo un análisis de equidad resulta de interés para poder evaluar si es posible aproximar a un análisis de equidad de un modelo de inteligencia artificial. Para ello, se toma un conjunto de datos denominado *German Credit* del repositorio online Kaggle⁸. Este contiene 1000 filas que representan a clientes de una organización bancaria y 10 columnas. De las columnas, 9 son atributos que contienen información característica de cada cliente y la restante es la clase objetivo a predecir. El objetivo del problema a resolver es predecir si un cliente entrará en *default* o no, es decir, se trata de una clasificación binaria (donde 1 representa riesgo de *default*, 0 riesgo de no *default*).

Tabla 2: variables del conjunto de datos

Variables	Descripción	Tipo de dato
Age	Edad	numérica
Sex	Género	categoría
Job	Ocupación	categoría
Housing	Si es propietario	categoría
Saving accounts	Tipo de ahorros	numérica
Checking account	Tipo de cuenta corriente	categoría
Credit amount	Monto del crédito	numérica
Purpose	Objetivo del crédito	numérica

Fuente: elaboración propia

Mediante la utilización del lenguaje de programación *Python* se realiza una preparación de los datos para luego aplicar un modelo de red neuronal secuencial. En primer lugar, las variables que son categóricas se convierten en variables *Dummies* (a excepción del género y la variable objetivo), es decir, cada valor de categoría pasa a ser una nueva variable. Cada una de esta toma valor 1 si el dato original tenía valor o 0 en caso contrario. De este modo, finalmente se obtienen 24 variables predictoras.

En los problemas de clasificación binaria, es frecuente que alguna de las categorías objetivo posea una cantidad de casos minoritaria (generalmente los casos identificados como riesgo de *default* o con valor 1). En el conjunto de datos a utilizar, se poseen 700 (70%) casos que son clientes sin

⁷ Acerca de Python <https://www.python.org/>

⁸ Kaggle es un sitio de competencias online donde diferentes organizaciones ponen a disposición conjuntos de datos para que los participantes logren encontrar el mejor modelo para resolver el problema planteado <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

riesgo de *default* y 300 (30%) con riesgo de *default*. Debido a que el desbalanceo en la clase objetivo suele afectar la performance de los modelos para realizar la predicción, se aplica un balanceo a través del método *Smote*⁹ y luego se estandarizan las variables. Finalmente, se divide el conjunto de datos en dos subconjuntos: uno para entrenar un modelo que contiene el 80% de los datos, y uno para testear la performance del modelo que contiene el 20% de los datos.

Con los datos ya preparados, a través de la librería Keras¹⁰ de *Python*, se configura la arquitectura de una red neuronal secuencial y se brindan diferentes opciones de valores de los parámetros de forma tal de encontrar la combinación óptima. Finalmente, como resultado de la optimización realizada, la arquitectura de la red aplicada contiene una capa de entrada con 32 neuronas, una capa intermedia con 16 neuronas y una capa de salida que contiene una única neurona.

Para evaluar la performance del modelo de inteligencia artificial como es el caso de la red neuronal para la clasificación de una clase binaria, las medidas utilizadas son el *Accuracy* y el Área bajo la Curva ROC (*Receiver Operating Characteristic Curve*, *AUC*). El *Accuracy* (exactitud) se define como el porcentaje de casos correctamente clasificados respecto del total de casos (Wang y Fu, 2005). Su cálculo viene dado por:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

donde TP, TN, FP y FN son los elementos contenidos en la matriz de confusión presentada en el apartado tercero.

El área bajo la Curva ROC es el área por debajo de la curva en un plano bidimensional (su cálculo es una integral) y brinda una medida agregada del rendimiento de todos los umbrales de clasificación posibles. La curva ROC compara la Sensibilidad con la Especificidad del modelo para diferentes umbrales. Cuanto más se acerca a uno mejor performance del modelo (Wang y Fu, 2005). La Sensibilidad y la Especificidad se definen como:

$$\text{Sensibilidad} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{Especificidad} = \frac{TN}{TN+FP} \quad (12)$$

donde TP, TN, FP y FN son los elementos contenidos en la matriz de confusión presentada en el apartado tercero.

⁹ Es un método para balancear la cantidad de casos por cada valor de la clase generando datos sintéticos. Para más acerca del método Smote ver <https://arxiv.org/pdf/1106.1813>

¹⁰ Acerca de Keras ver <https://keras.io/>

Obteniéndose las predicciones del modelo optimizado aplicado sobre el subconjunto de datos de testeo y utilizando la librería *Scikit-Learn*¹¹ para el cálculo de las medidas de performance presentadas anteriormente, se obtiene un *Accuracy* del 0.796 y un AUC de 0.790. En base a estos resultados, es posible decir que el modelo de red neuronal tiene una exactitud de predicción aceptable para casos correctos en base a los datos utilizados.

Sin embargo, entre las variables predictoras se detecta una variable sensible que es el género. Entre los datos utilizados, resulta que hay 310 clientes que son de género femenino y 690 de género masculino, pudiéndose interpretar la existencia de un posible sesgo hacia este último género al ser un grupo mayoritario. Frente a esta situación, resulta necesario evaluar si el modelo está siendo equitativo frente a la disparidad de casos por género. Para ello, a continuación se aplica la metodología presentada en el apartado anterior para su análisis.

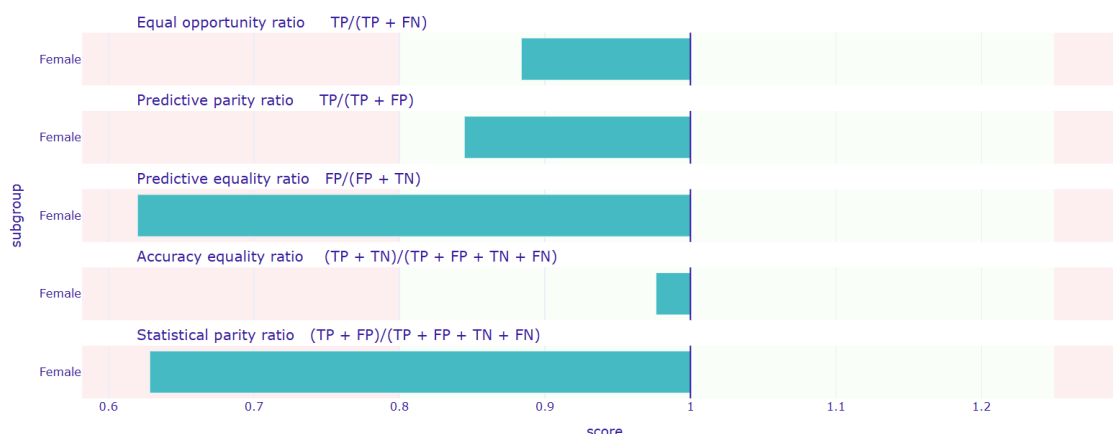
4.1 ANÁLISIS DE EQUIDAD SEGÚN GÉNERO

En base a las predicciones obtenidas sobre el subconjunto de datos de testeo luego de haber aplicado el modelo de red neuronal pre entrenado, se realiza un análisis de equidad por género con *Python*. Para ello se utiliza la librería *Dalex*¹², que calcula las medidas más relevantes al problema de equidad definidas en el apartado tercero. Como resultados se obtienen los siguientes valores de cada una de las medidas, las cuales son visualizadas del siguiente modo:

¹¹ Acerca de Scikit-Learn ver <https://scikit-learn.org/stable/>

¹² Acerca de *Dalex* <https://dalex.drmhy.ai/>

Gráfico 1: medidas de equidad iniciales con datos sin balancear por género



Fuente: Elaboración propia con *Python*

En el gráfico 1, las barras verdes muestran el valor obtenido de cada medida visualizada en términos de *score*, como se explicó en el apartado tercero. Además, se visualiza en el fondo de este, una zona identificada con color rojizo y otra verdoso. Cuando un *score* de las medidas supera la zona verdosa, es un indicador de que el modelo está siendo injusto para el grupo minoritario frente al mayoritario. Dado que al menos dos medidas se encuentran bajo dicha situación, es posible concluir que el modelo es inequitativo frente al grupo minoritario que son los clientes de género femenino.

Con el fin de especificar el valor del *score* obtenido en cada caso, a continuación, se muestra el valor de cada medida por cada grupo; y en la última fila el cociente entre ambas siendo el *score* visualizado:

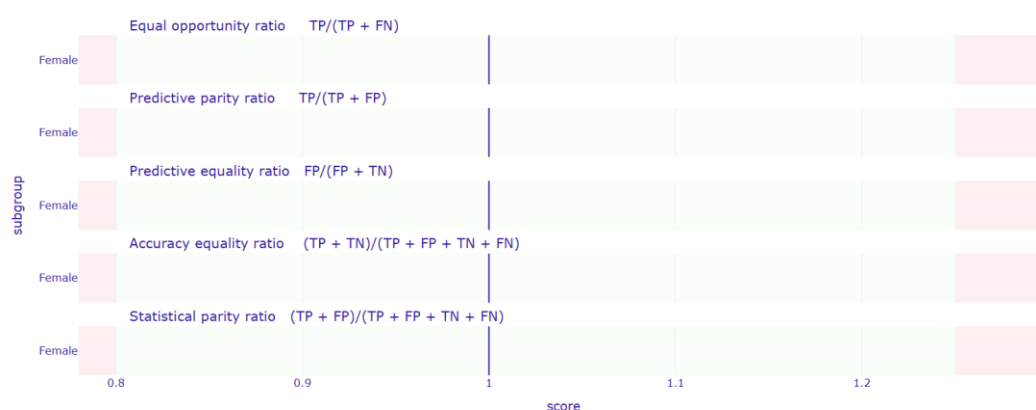
Tabla 2: medidas absolutas de equidad y obtención de score

	Equal opportunity ratio	Predictive parity ratio	Predictive equality ratio	Accuracy equality Ratio	Statistical parity ratio
Female	0.230	0.806	0.230	0.784	0.443
Male	0.371	0.912	0.371	0.803	0.705
Cociente=Score	0.619	0.883	0.619	0.976	0.628

Fuente: Elaboración propia con *Python*

De este modo, es posible conocer si un modelo está replicando un sesgo humano que se encuentra reflejado en los propios datos. A partir de ello se habilita la posibilidad de tomar una decisión para corregir dicho problema de forma tal de tener un accionar responsable. Entre diversas acciones posible, se encuentra el poder balancear los datos mediante *Smote*, pero ahora en base a la variable sensible que es el género. Llevándolo a cabo con *Python*, volviendo a ejecutar el modelo y posteriormente repitiendo el análisis de equidad, se obtiene:

Gráfico 2: medidas de equidad post balanceo por género



Fuente: Elaboración propia con *Python*

En el gráfico 2 es posible notar que ahora ninguna de las medidas supera el umbral aceptable de 0.8. Por el contrario, todas ellas valen 1 o muy cercanas a este valor. De esta manera, se pudo corregir la inequidad presentada en los datos y en consecuencia el modelo aplicado ya no reproduce tal discriminación por grupo sensible.

CONCLUSION

A lo largo del presente trabajo se ha mostrado que es posible detectar si un modelo de inteligencia artificial reproduce un sesgo humano de discriminación. La utilización de variables sensibles, como el género, para realizar predicciones con un modelo de red neuronal, muestra que estos modelos no detectan dicho sesgo, sino que lo reproducen. Frente a ello resulta necesario abordar medidas analíticas que eviten la reproducción de este problema de forma tal de contribuir a un uso responsable de la AI. De este modo, se podrá facilitar el acceso a mitigar diferentes riesgos asociados al uso de estas tecnologías para actuar bajo un marco de responsabilidad.

Para abordar la detección de discriminación de un modelo de IA en base a alguna variable sensible, resulta necesario aplicar alguna metodología cuantitativa que, de manera clara y sencilla, facilite dicha tarea. En base al ejemplo desarrollado en el presente trabajo, la metodología propuesta para el análisis de equidad resultó efectiva. Se logró detectar que el modelo de red neuronal aplicado a los datos utilizados reproducía un sesgo discriminatorio por género.

Una vez que el problema de inequidad por género fue detectado, fue posible subsanarlo y continuar utilizando el modelo de IA. Lo importante aquí es que se puso de relieve que es necesario que quien toma la decisión – es decir, quien utiliza el modelo- debe procurar entrenar dichos modelos con una preparación de datos adecuada. De este modo, se podrá evitar que el modelo reproduzca un sesgo humano, lo que delimita un accionar bajo un marco de responsabilidad.

Si bien lo realizado en este trabajo se limita a un caso de aplicación, se considera que constituye un primer paso para contribuir a un uso responsable de la inteligencia artificial. En este sentido, resulta de interés llevar adelante una ampliación de lo realizado. Entre diversas posibilidades, sería adecuado realizar la aplicación para modelos de regresión de forma tal de conocer si la metodología de propuesta resulta igualmente eficiente.

BIBLIOGRAFÍA DE REFERENCIA

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, 99, 101805. DOI: <https://doi.org/10.1016/j.inffus.2023.101805>

Banco Central de la República Argentina (2023). Normativa A7724. Disponible en: <https://www.bcra.gob.ar/pdfs/comytexord/A7724.pdf>

Chen, Z., Xiao, F., Guo, F. & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*, Volume 9, 2023, 100123. ISSN 2666-7924. DOI: <https://doi.org/10.1016/j.adapen.2023.100123>.

Brzezinski, D., Stachowiak, J., Stefanowski, J., Szczech, I., Susmaga, R., Aksenyuk, S. & Yasinskyi, O. (2024). Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, 18(7), 1-18. DOI: <https://doi.org/10.1145/3654659>

Dignum, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 2156). Cham: Springer. DOI: <https://doi.org/10.1007/978-3-030-30371-6>

Disposición 2/2023. “Recomendaciones para una Inteligencia Artificial Fiable”. <https://www.boletinoficial.gob.ar/detalleAviso/primera/287679/20230602>

Erickson, J. (1999). Algorithms. Creative Commons Attribution 4.0 International license. <https://jeffe.cs.illinois.edu/teaching/algorithms/book/Algorithms-JeffE.pdf>

Hromkovič, J. (2005). Fundamentals. In Hromkovič, J, Design and Analysis of Randomized Algorithms. Introduction to Design Paradigms (pp. 19-99). Springer Berlin, Heidelberg. ISBN: 978-3-540-27903-7.

Kolanovic, M., & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. JP Morgan Global Quantitative & Derivatives Strategy Report. <https://cpb-us-e2.wpmucdn.com/faculty.sites.uci.edu/dist/2/51/files/2018/05/JPM-2017-MachineLearningInvestments.pdf>

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1-35. DOI: <https://doi.org/10.1145/3457607>

OECD, (2019). “AI Principles overview”. <https://oecd.ai/en/ai-principles>

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3), 1-44. DOI: <https://doi.org/10.1145/3494672>

UNESCO, (2022). “Recomendación sobre la ética de la inteligencia artificial”.
https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Wang, L., & Fu, X. (2005). Data mining with computational intelligence. Springer Science & Business Media. DOI: https://doi.org/10.1007/3-540-28803-1_3