Mundo Fesc E-ISSN 2216-0353 P-ISSN 2216-0388P-ISSN 2216-0388 **Artículo Original**

Equidad en aprendizaje automático: comparación de intervenciones de preprocesamiento y entrenamiento en un caso colombiano

Fairness in machine learning: A comparison of preprocessing and in-processing interventions in a Colombian case study

Recibido: 12 de febrero de 2026
Aprobado: 16 de julio de 2026

Forma de citar: A. . Rosado Gomez, "Equidad en aprendizaje automático: comparación de intervenciones de preprocesamiento y entrenamiento en un caso colombiano", Mundo Fesc, vol. 16, no. 36, pp. 94- 125 Sep. 2026, doi: 10.61799/2216-0388.2338.

Alveiro Rosado Gomez
Ph.D en Ingeniería
aarosadog@ufpso.edu.co
https://orcid.org/0000-0003-2932-3383
Universidad Francisco de Paula Santander
Ocaña, Colombia.

*Autor para correspondencia:
aarosadog@ufpso.edu.co

© 2026. Fundación de Estudios Superiores Comfanorte.

Equidad en aprendizaje automático: comparación de intervenciones de preprocesamiento y entrenamiento en un caso colombiano

Resumen

Esta investigación tuvo como objetivo evaluar si la equidad alcanzada por clasificadores tradicionales mediante intervenciones de preprocesamiento puede transferirse a un perceptrón multicapa mediante destilación de conocimiento, y si el uso del regularizador MinDiff aporta una mejora adicional en la reducción de la disparidad de falsos negativos entre sexos en un clasificador de acceso al Beneficio de Inserción Económica en Colombia. Para ello se aplicó un diseño experimental por fases con registros administrativos de personas desmovilizadas: con una arquitectura de perceptrón multicapa fija se evaluaron diferentes intervenciones que combinaron reponderación, análisis de componentes principales, MinDiff, adversarial debiasing y seis estrategias de destilación desde dos maestros mitigados, con cinco semillas, tres particiones estratificadas, intervalos de confianza BCa, prueba de Wilcoxon con corrección de Holm y frontera de Pareto. Los experimentos mostraron que la destilación de los modelos maestros previamente analizados produjo una disminución en la diferencia entre tasas de falsos negativos, en comparación con el perceptrón multicapa de referencia. La destilación redujo la diferencia entre las tasas de falsos negativos respecto al MLP de referencia. Los menores valores se obtuvieron al combinar las probabilidades de los maestros. Con ponderación por exactitud y MinDiff, FNR_DIF pasó de 0.000982 a 0.000502, mientras el recall se mantuvo en 0,999598. La reducción de FNR_DIF fue descriptiva entre las ejecuciones; en DPD y recall sí se obtuvieron diferencias estadísticamente significativas. Accuracy presentó una disminución moderada.

Palabras clave: aprendizaje automático; destilación de conocimiento; equidad algorítmica; inteligencia artificial; toma de decisiones.

Fairness in machine learning: A comparison of preprocessing and in-processing interventions in a Colombian case study

Abstract

This paper aimed to evaluate whether the fairness achieved by traditional classifiers through preprocessing interventions could be transferred to a multilayer perceptron via knowledge distillation, and whether the MinDiff regularizer provided further improvement in reducing the false-negative rate disparity between genders in a classifier determining eligibility for the Economic Insertion Benefit in Colombia. A phased experimental design using administrative records of demobilized individuals was employed: using a fixed multilayer perceptron architecture, the study evaluated various interventions combining reweighting, principal component analysis, MinDiff, adversarial debiasing, and six distillation strategies from two mitigated teacher models. The evaluation involved five random seeds, three stratified data splits, BCa confidence intervals, the Wilcoxon test with Holm correction, and Pareto frontier analysis. The experiments demonstrated that distilling the previously analyzed teacher models resulted in a decrease in the false-negative rate disparity compared to the baseline multilayer perceptron. Distillation reduced the gap in false-negative rates relative to the baseline MLP. The lowest values were achieved by combining the teachers' probabilities; with accuracy-based weighting and MinDiff, the FNR disparity (FNR_DIF) dropped from 0.000982 to 0.000502, while recall remained at 0.999598. The reduction in FNR_DIF was observed across runs, whereas statistically significant differences were found regarding DPD and recall. Accuracy showed a moderate decrease.

Keywords: algorithmic fairness; artificial intelligence; decision making; knowledge distillation; machine learning.

Introducción

El auge del aprendizaje automático ha permitido automatizar decisiones en campos como el empleo, la salud o el acceso a créditos; sin embargo, aprender de datos históricos puede consolidar o amplificar sesgos preexistentes [1] [2]. Por ello, los sesgos deben considerarse en todas las fases del ciclo de construcción de aplicaciones de apoyo a la decisión, supervisando los datos desde su recolección hasta el momento en que se emite una recomendación [3] [4] [5].

El sesgo surge principalmente de conjuntos de datos que contienen distorsiones generadas por la acción humana [6] [7], relacionadas con la influencia de atributos protegidos o sensibles, como el sexo, la raza, la nacionalidad o la religión, sobre la clasificación [1]. Según los valores de dichos atributos, las personas pertenecen a grupos privilegiados o no privilegiados, y las salidas del modelo resultan en general más favorables para los primeros [3] [4].

Las soluciones a este problema constituyen un campo de investigación ética y algorítmica orientado a comprender cómo se manifiesta el fenómeno y a desarrollar mecanismos de control y herramientas que aseguren decisiones equitativas [8] [9] [10]. Sin embargo, aún no existe un marco común y estandarizado con una secuencia de pasos clara para mitigar el sesgo, por lo que se requieren iniciativas que aporten componentes para estructurar su solución [11].

Un comportamiento común de las soluciones consiste en establecer una línea base con métricas de rendimiento y equidad y compararla con los resultados tras las intervenciones; los mejores resultados disminuyen la discriminación sin comprometer el rendimiento [11] [12] [13] [14]. La propuestas de intervenciones están asociadas a las fases de construcción de los modelos: el preprocesamiento modifica los datos antes del entrenamiento; el procesamiento (in-processing) introduce la equidad dentro del aprendizaje mediante regularizadores o restricciones; y el posprocesamiento ajusta las salidas del modelo entrenado [5] [3] [4] [12]. Esta investigación utiliza como referente principal los resultados de un estudio previo sobre datos de personas desmovilizadas de grupos armados ilegales del conflicto armado en Colombia y que siguieron el proceso de reintegración a la sociedad. En dicha investigación se planteó el caso de uso de la construcción de un clasificador binario que, bajo criterios de equidad, determino los desmovilizados que deberían acceder al Beneficio de Inserción Económica (BIE) [15].

Este interrogante nace a partir de un estudio que reportó que las intervenciones de preprocesamiento reducen los falsos negativos (False negative rates, FNR) sin eliminar el compromiso entre desempeño y equidad: un AdaBoostClassifier (ADA) optimizado por ajuste de hiper-parámetros alcanzó una diferencia de FNR baja, y un RandomForestClassifier (RFC) con reponderación y selección de características derivada de análisis de componentes principales (Principal component analysis, PCA) mostró resultados compensados entre desempeño y robustez. Todo ello, a pesar de que la mayoría de los individuos eran hombres, y estos resultaron ser el grupo no privilegiado, porque los clasificadores cometían más errores de asignación del beneficio con ellos que con las mujeres [15].

El estudio referente mostró que las intervenciones de preprocesamiento reducen los falsos negativos (False negative rates, FNR) manteniendo el compromiso entre desempeño y equidad: un AdaBoostClassifier (ADA) optimizado por ajuste de hiper-parámetros alcanzó una diferencia de FNR baja, y un RandomForestClassifier (RFC) con reponderación y selección de características derivada de análisis de componentes principales (Principal component analysis, PCA) mostró resultados compensados entre desempeño y robustez [15].

Partiendo de la premisa reportada en la investigación de antecedente, se plantea la pregunta: ¿puede el comportamiento aprendido por un modelo transferirse a una red neuronal mediante destilación de conocimiento?

La destilación de conocimiento fue introducida como una estrategia para entrenar modelos estudiantes que imitan las salidas suaves de modelos maestros, transfiriendo información no contenida en la etiqueta dura de clase [16] [17]. Sin embargo, si el maestro no es equitativo o si la combinación de maestros no se controla, puede reproducir o amplificar sesgos [18] [19], por lo tanto, dicha posibilidad exige una evaluación empírica.

Dentro de las intervenciones durante el procesamiento, se encuentran soluciones que introducen penalidades para incluir la equidad. MinDiff, es una de ellas y tiene como finalidad aplicar un criterio de equidad mediante una penalización basada en la discrepancia de máxima media para acercar las distribuciones de salida entre grupos, dentro de la familia de intervenciones durante el procesamiento [20] [21] [22]; su efectividad, puede depender de la capacidad del modelo, de la forma de regularización y de la relación entre el objetivo predictivo y el atributo sensible [23].

El vacío que aborda este trabajo está en la conexión entre dos momentos de mitigación que la literatura suele comparar como familias separadas [5] [24]. La aplicación sobre destilación para la equidad se viene trabajando enfocado en las redes neuronales, concentrado en visión por computador, lenguaje natural y aprendizaje federado [18] [25]; la intersección entre destilación desde clasificadores tradicionales no neuronales, datos tabulares y equidad de grupo permanece poco explorada [26]. La contribución consiste en evaluar, sobre un caso real colombiano, la transferencia de equidad desde clasificadores tradicionales preprocesados hacia modelos neuronales mediante destilación, analizando además la contribución de intervenciones hechas especialmente durante el entrenamiento mediante MinDiff al equilibrio entre exactitud, sensibilidad y disparidad de falsos negativos. El estudio se guía por tres preguntas de investigación: si un perceptrón multicapa (Multilayer perceptrón, MLP) destilado puede preservar o mejorar la equidad de sus maestros; qué estrategia de combinación de maestros produce mejores resultados; y si MinDiff mejora la equidad o introduce un costo innecesario cuando se combina con destilación. El resto del artículo presenta la fundamentación teórica, el método por fases, los resultados de las fases sin y con destilación, la discusión frente a la literatura y las conclusiones.

Equidad, sesgo y atributos protegidos

La equidad en aprendizaje automático se refiere a la ausencia de tratamiento injusto hacia individuos o grupos en la predicción de un modelo [2]. Los conjuntos de datos incorporan decisiones humanas pasadas y patrones de discriminación que el modelo puede aprender [7] [27]. Existen dos alcances de la equidad: la equidad individual, que exige tratar de forma similar a individuos similares, y equidad de grupo, que compara tasas de error o de asignación entre subpoblaciones definidas por un atributo protegido [8]. Este trabajo se sitúa en la equidad del atributo sensible sexo, bajo una variante de la igualdad de oportunidades: que la probabilidad de negar erróneamente el beneficio a una persona elegible no dependa de su sexo.

La paridad demográfica exige tasas de asignación positiva similares entre grupos, mientras que la igualdad de oportunidades exige tasas de verdaderos positivos similares [2] [3]. Cuando la clase positiva domina el conjunto de datos, como aquí, un modelo puede exhibir buena paridad demográfica y concentrar sus falsos negativos en un grupo, lo que motiva usar la diferencia de tasas de falsos negativos como métrica de equidad [13] [12].

Métricas de desempeño y de equidad

Las métricas de equidad de grupo se derivan en su mayoría de la matriz de confusión calculada por subgrupo y evalúan la diferencia o proporción de la discriminación entre los subgrupos observados [2]. Entre las métricas de desempeño, la exactitud (accuracy, ACC) es la fracción de predicciones correctas; la sensibilidad (recall, REC) mide el porcentaje de positivos identificados correctamente; y la tasa de falsos negativos (FNR) es el cociente entre positivos mal clasificados y el total de positivos del grupo. Siguiendo lo establecido en el artículo base de esta investigación, se tomó como medida primaria de equidad la diferencia absoluta de FNR entre hombres y mujeres (FNR_DIF), complementada con la diferencia de paridad demográfica (DPD), que compara las tasas de asignación positiva entre grupos, y el impacto dispar (DI), que expresa la razón entre dichas tasas; valores de DPD cercanos a cero y de DI cercanos a uno indican menor disparidad [3].

La elección de FNR_DIF como métrica primaria responde al contexto de aplicación del modelo: el error más costoso es negar el beneficio a una persona que culmino su ruta de reinserción. Por ello el desempeño se evaluó en este dominio, que la probabilidad de reconocer a una persona elegible no dependa de su sexo, midiendo la distribución del error entre grupos. El error más grave es negar el beneficio a una persona elegible y, en un conjunto dominado por la clase positiva, la disparidad entre grupos se concentra precisamente en los falsos negativos [2] [28].

Familias de intervenciones de mitigación

Las intervenciones de mitigación se organizan según el momento del ciclo de modelado en el que actúan [5] [3]. La Tabla 1, sintetiza las tres familias y ubica la contribución del estudio: conectar el preprocesamiento, materializado en los maestros del estudio previo, con la familia de intervenciones aplicada durante el entrenamiento, tanto del estudiante mediante la destilación de conocimiento y con MinDiff.

Tabla 1. Comparación entre familias de intervenciones de mitigación de sesgo

Familia	Objetivo	Técnicas representativas	Ventajas	Limitaciones
Preprocesamiento	Modificar datos y atributos antes del entrenamiento para alterar la relación entre atributo sensible y etiqueta.	Reponderación, remuestreo, supresión del atributo, representaciones justas, PCA.	Independiente del clasificador, fácil de integrar en pipelines existentes.	No controla directamente el criterio de equidad del modelo final; puede degradar información.
Procesamiento (in-processing)	Incorporar la equidad en la función objetivo del aprendizaje.	Regularizadores (MinDiff/MMD), restricciones, adversarial debiasing, desfijación con maestros mitigados.	Control explícito del criterio de equidad; ajusta la frontera de decisión.	Depende de la capacidad del modelo y de la intensidad de la regularización; costo potencial en desempeño.
Posprocesamiento	Ajustar las salidas del modelo entrenado para satisfacer un criterio de equidad.	Umbrales por grupo, calibración, reordenamiento de puntuaciones.	No requiere reentrenar; aplicable a modelos de caja negra.	Puede requerir el atributo sensible en inferencia; ajusta salidas y no causas.

Intervenciones de preprocesamiento empleadas por los modelos maestros

En el estudio original se aplicaron técnicas de preprocesamiento y optimización como reponderación, supresión del atributo sensible, balanceo, PCA y optimización de hiper-parámetros [15]. La reponderación asigna pesos a las instancias para que la distribución ponderada resulte independiente del atributo protegido [29]. El PCA reduce la dimensionalidad y puede disminuir el peso de variables correlacionadas con patrones discriminatorios, aunque no garantiza equidad por sí mismo; existen variantes diseñadas explícitamente para reducir la codificación demográfica de las representaciones [30]. Siguiendo las pautas dadas por [31], tanto PCA como la optimización de hiper-parámetros pueden considerarse intervenciones para disminuir la discriminación durante el preprocesamiento, produciendo efecto positivo sobre métricas de grupo [32] [33].

Del estudio mencionado se seleccionaron las dos combinaciones más eficaces y robustas: un AdaBoostClassifier optimizado mediante optimización de hiper-parámetros sobre los datos originales (ABC+PHT) y un RandomForestClassifier entrenado con reponderación y selección de columnas proveniente del pipeline de PCA (RFC+REW+PCA) [15]. Estos dos modelos fueron utilizados como maestros para la destilación, representando perfiles complementarios de paridad y exactitud.

MinDiff y discrepancia de máxima media

MinDiff es una técnica de regularización durante el entrenamiento que añade al objetivo de aprendizaje una penalización basada en la discrepancia de máxima media (Maximum Mean Discrepancy, MMD) para acercar las distribuciones de salidas condicionadas por el grupo [21] [34] [35]; para lograr una medida más fuerte de equidad, la correlación entre predicciones y atributo sensible se reemplaza por la MMD entre las distribuciones de puntuaciones de ambos grupos [21].

Empíricamente, MinDiff puede reducir la brecha de falsos negativos con una pérdida mínima de exactitud, aunque en modelos que interpolan el conjunto de entrenamiento la penalización pierde efectividad salvo que se combine con regularización adicional [34] [23] [35]. Esto motiva dos decisiones de diseño: mantener fija la capacidad del estudiante y recalibrar la intensidad de la regularización cuando cambia la señal de aprendizaje.

Destilación de conocimiento y su relación con la equidad

La destilación de conocimiento (knowledge distillation, KD) transfiere el comportamiento de un modelo maestro hacia un modelo estudiante, de modo que este alcance un rendimiento comparable [16] [18]; históricamente se viene utilizando como una técnica de comprensión de modelos [36]. En el esquema maestro-estudiante, el maestro genera una señal más rica que las etiquetas duras, típicamente probabilidades suavizadas que expresan la incertidumbre relativa y la frontera de decisión, y el estudiante combina una pérdida de destilación con la pérdida supervisada clásica [17] [37].

Esta forma de aprendizaje permite además de la generalización, encaminar el aprendizaje a reducir el modelo destilado [18] [38]. De esta forma se pueden tener maestros entrenados solo para exactitud con restricciones de equidad en el estudiante [39]; o maestros sesgados especializados por grupo que se combinan para guiar a un estudiante para reducir disparidades [40] [19].

La forma de configurar los maestros puede influir en el resultado final del estudiante [26], por ejemplo haciendo aprendizaje incremental en donde el sesgo se concentra en la última capa [38], otro factor que puede influir es que la condensación de conjuntos de datos tiende a empeorar la inequidad hacia grupos minoritarios salvo que se estructure por grupos protegidos [26]; también existe la posibilidad que los estudiantes pueden heredar parcialmente la equidad de sus maestros [41], sin embargo, la destilación ofrece una optimización más estable que la optimización directa de la equidad. La literatura, también señala que la equidad se incorpora deliberadamente al proceso, y ambivalente cuando la destilación se aplica de forma ingenua [18] [39] [26] [25] [37]. Este balance justifica esta investigación: maestros mitigados, mediante la combinación con reglas determinísticas y verificación estadística de la transferencia.

Comparación multiobjetivo y análisis estadístico

Dado que se busca lograr la mejor combinación que genere una compensación entre rendimiento y equidad, el resultado no es un modelo único, sino un grupo de soluciones, se traslada la decisión final al tomador de decisiones, para lograrlo se utiliza el frente de Pareto, para analizar el conjunto de configuraciones no dominadas en el que mejorar un objetivo implica empeorar otro, dado que no todas las intervenciones equilibran equidad y rendimiento [3] [5].

Asimismo, para evitar declarar diferencias significativas por azar cuando se contrastan muchas hipótesis [42], se requieren ajustes, y dado que se comparan múltiples intervenciones bajo condiciones experimentales repetidas, se necesitan métodos estadísticos que distingan si las diferencias observadas son consistentes o dependen de la variabilidad de semillas y particiones [43].

El procedimiento de Holm [44], fue utilizado para hacer múltiples comparaciones frente a una línea base común, controlando la tasa de error familiar, además de los tamaños del efecto, porque el valor p no mide magnitud [45]; el coeficiente rank-biserial expresa la diferencia entre evidencia favorable y desfavorable [46], mientras que la prueba de rangos con signo de Wilcoxon [47] compara condiciones relacionadas sin exigir normalidad y es la recomendada para clasificadores con resultados pareados [43], dado que las métricas de clasificación y equidad suelen estar presentadas en distribuciones asimétricas.

Diseño y contexto del estudio

Se establecieron un conjunto de fases para evaluar si la equidad alcanzada por clasificadores tradicionales mediante preprocesamiento puede transferirse a un MLP mediante destilación de conocimiento, y si MinDiff aporta una mejora adicional. El caso de uso es la clasificación binaria del acceso al Beneficio de Inserción Económica (BIE) de personas desmovilizadas del conflicto armado en Colombia; la clase positiva representa el acceso al beneficio y el atributo sensible fue el sexo, codificado de forma binaria (grupo g0, mujeres; grupo g1, hombres). La métrica primaria fue FNR_DIF porque el error más costoso del dominio es el falso negativo: negar el beneficio a una persona elegible.

Población, muestra y criterios de selección

El conjunto inicial descargado del portal de Datos Abiertos del Gobierno de Colombia contenía 55655 registros y 37 columnas. Se conservaron únicamente las personas que ingresaron al proceso de reintegración y que contaban con censo de familia. Se eliminaron columnas por: constantes, otras que tenian cerca del 50 % de valores <No Registra>. Se excluyeron luego los registros con valores <No Aplica>, <No Registra> o inválidos; los valores restantes se imputaron. En la tabla 2, se muestran los atributos que se tuvieron en cuenta tanto en el estudio original [15], como en la presente investigación.

Tabla 2. Atributos utilizados en la investigación

Variable original	Tipo
Desembolso BIE (etiqueta)	Categórica binaria
Sexo (atributo sensible)	Categórica binaria
Año de desmovilización	Numérica
Año de independización/ingreso	Numérica
Grupo etario	Ordinal (4)
Departamento de residencia	Categórica (alta cardinalidad)
Municipio de residencia	Categórica (alta cardinalidad)
Beneficio TRV / FA / FPT / PDT (4 variables)	Binarias
Nivel educativo	Ordinal (4)
Ocupación económica	Categórica (5)
Posee servicio social	Ordinal (3)
Posee cónyuge o compañero(a)	Categórica binaria
Número de hijos	Numérica
Total integrantes del grupo familiar	Numérica
Tipo de vivienda	Categórica (8 tras agrupar)
Régimen de tenencia de la vivienda	Categórica
Posee servicios públicos básicos	Categórica binaria
Régimen de salud	Categórica

La Tabla 3, muestra la distribución de los subconjuntos según las combinaciones de sexo, femenino (S=F) o masculino (S=M), con la clase (C=0 o C=1), y evidencia el doble desbalance del problema, por clase y por grupo. En entrenamiento, las combinaciones S=M,C=1; S=M,C=0; S=F,C=1 y S=F,C=0 representan el 64.8 %, 16.7 %, 16.0 % y 2.5 % de los registros, y en prueba el 65.2 %, 16.3 %, 16.7 % y 1.8 %; la clase positiva concentra cerca del 81 % de los casos y el grupo masculino cerca del 81.5 %, lo que configura el doble desbalance, por clase y por grupo.

Tabla 3. Distribución de los subconjuntos de datos

Subconjunto	S=M, C=1	S=M, C=0	S=F, C=1	S=F, C=0	Total
Entrenamiento	11071	2856	2742	425	17094
Prueba	2787	695	715	77	4274

Del estudio original se utilizaron ABC+PHT y RFC+REW+PCA, los dos clasificadores con mejor equilibrio entre desempeño, equidad y robustez. Como referencia externa, los valores publicados de esos modelos reportan para ABC+PHT una exactitud de 0.8200, una FNR de 0.0004 en hombres y de 0.0000 en mujeres y un recall de 1.0000, y para RFC+REW+PCA una exactitud de 0.8000, una FNR de 0.0477 en hombres y de 0.0460 en mujeres y un recall de 0.9500 [15].

La arquitectura propuesta para llevar el estudio base a un perceptrón multicapa fue tomado mediante la arquitectura que fue con formada por 54 configuraciones de rejilla, la intensidad lambda de MinDiff sobre {0.1, 0.5, 1.0, 2.0, 5.0}, el coeficiente alpha de destilación sobre {0.3, 0.5, 0.7} y la lambda de adversarial debiasing, se tomaron utilizando exclusivamente particiones internas de entrenamiento/validaciones derivadas del 80 % de entrenamiento, con las tres primeras semillas (13, 17 y 23). El conjunto de prueba no participó en ninguna decisión de calibración, ajuste o selección de configuraciones y se utilizó una única vez por ejecución en la evaluación final.

Variables, categorías e instrumentos

Las expresiones se presentan en formato lineal. La FNR por grupo sensible g y la métrica primaria se definen en (1) y (2); el peso de reponderación por grupo s y clase y, con n el total de instancias, ns y ny los marginales y nsy el conteo conjunto, se define en (3) [29].

$$FNR_g = \frac{FN_g}{(FN_g + TP_g)} \quad (1)$$

$$FNR_{DIF} = \left| FNR_{g0} - FNR_{g1} \right| \quad (2)$$

$$w_{(s,y)} = \frac{(n_s * n_y)}{(n * n_{sy})} \quad (3)$$

MinDiff fue implementado como un regularizador en tiempo de entrenamiento propio en TensorFlow. En cada lote de entrenamiento se toman submuestras balanceadas de los casos positivos de cada grupo sensible (de tamaño igual a un cuarto del lote por grupo, exigiendo al menos dos positivos por grupo) y se calcula la discrepancia de máxima media entre las distribuciones de probabilidades predichas de ambos subconjuntos, con el estimador (ecuación 4 y 5) entre las predicciones de los grupos. La penalización se suma a la entropía cruzada binaria multiplicada por el coeficiente lambda (ecuación 6), de modo que el gradiente de equidad actúa en cada paso de optimización.

$$k_{(a,b)} = exp\left(\frac{-\|a-b\|^2}{2*sigma^2}\right) \quad (4)$$

$$MMD^2(P0,P1) = mean(k(P0,P0)) - 2 * mean(k(P0,P1)) + mean(P1,P1)) \quad (5)$$

$$L_{noKD_{MD}} = L_{CE} + lambda * L_{MD} \quad (6)$$

Para la destilación, la etiqueta suave combinada se obtiene como mezcla convexa de las probabilidades de los maestros (ecuación 7). La pérdida de destilación es la divergencia KL binaria con temperatura T entre la etiqueta suave del maestro (tT) y la predicción del estudiante (sT), escalada por T2 (ecuación 8) [16]. La pérdida total combina señal supervisada, señal destilada y MinDiff (ecuación 9); alpha controla el peso de la etiqueta real (alpha igual a 0.3 significa 30 % de pérdida supervisada y 70 % de destilación) y en las variantes sin MinDiff se fija lambda igual a cero.

$$p_{KD(x)} = w_{ABC} * p_{ABC(x)} + w_{RFC} * p_{RFC(x)} \quad (7)$$

$$L_{KD} = \text{mean}\left[t_T * \log\left(\frac{t_T}{s_T}\right) + (1 - t_T) * \log\left(\frac{(1 - t_T)}{(1 - s_T)}\right)\right] * T^2 \quad (8)$$

$$L_{total} = alpha * L_{CE} + (1 - alpha) * L_{KD} + lambda * L_{MD} \quad (9)$$

Procedimiento y análisis de datos

Secuencia metodológica del proceso

El flujo neuronal se implementó en Python 3.12.13 con TensorFlow 2.20.0 y Keras 3.13.2, usando scikit-learn para la estandarización, el PCA y las particiones estratificadas, y SciPy para las pruebas estadísticas y el remuestreo bootstrap. Los maestros se ejecutaron en un contenedor Docker compatible con las versiones originales de sus bibliotecas, desde el cual se exportaron sus probabilidades suaves y métricas operativas como archivos planos consumidos por el flujo neuronal, y todos los procesos aleatorios se controlaron con semillas explícitas. El estudio se organiza en dos fases experimentales, una sin destilación y otra con destilación. En la Tabla 4 se muestra la metodología que siguió la investigación, mediante fases que estructuran las reglas, configuraciones y resultados reproducibles del flujo final.

Tabla 4. Secuencia metodológica del estudio.

Fase	Nombre	Propósito	Salida metodológica
Fase 0	Estudio original	Identificar modelos tradicionales con buen equilibrio entre rendimiento y equidad.	ABC-PHT y RFC+REW+PCA como maestros candidatos.
Fase 1	Calibración del MLP	Buscar una arquitectura neuronal base antes de aplicar mitigaciones.	Arquitectura (64, 32), dropout 0.5, L2 0.0, tasa de aprendizaje 5×10^{-4}.
Fase 2	Intervenciones sin desilación	Evaluar REW, PCA, MinDiff y sus combinaciones sobre el MLP fijo.	Lambda de MinDiff igual a 5.0 en esta fase.
Fase 3	Etiquetas suaves	Generar probabilidades de los maestros sin reentrenarlos en el flujo neuronal.	$P_{ABC}(X)$ y $P_{RFC}(X)$.
Fase 4	Cálculo de pesos KD	Combinar maestros por promedio, exactitud, equidad y criterio multiobjetivo.	Pesos determinísticos, no aleatorios.
Fase 5	Destilación con y sin MinDiff	Entrenar estudiantes MLP con destilación, con y sin MinDiff.	Alpha 0.3 y lambda de MinDiff igual a 0.1 en esta fase.
Fase 6	Evaluación final	Comparar intervenciones con IC BCa, Wilcoxon-Holm y frontera de Pareto.	Veinte intervenciones, cinco semillas, tres particiones.

Calibración y fijación de la arquitectura MLP

Antes de aplicar las intervenciones se fijó una única configuración del MLP. Para seleccionarla se probaron 54 combinaciones de arquitectura, dropout, regularización L2 y tasa de aprendizaje. La selección se realizó tomando FNR_DIF como primer criterio; cuando dos configuraciones presentaban resultados similares, se compararon recall y accuracy. La configuración seleccionada utilizó dos capas ocultas de 64 y 32 neuronas, activación ReLU, dropout de 0,5, sin regularización L2 y Adam con una tasa de aprendizaje de 5×10^{-4}. Esta arquitectura se mantuvo sin cambios durante los experimentos posteriores [23]. La Tabla 5 resume la configuración y las decisiones de entrenamiento y de control de fuga de información aplicadas.

Tabla 5. Configuración base del MLP y decisiones de entrenamiento

Elemento	Valor o regla
Capas ocultas	(64, 32), activación ReLU
Dropout / L2 / tasa de aprendizaje	0.5 / 0.0 / 5×10^{-4} (Adam)
Salida y umbral	Sigmoide de una neurona; umbral fijo de 0.5
Épocas máximas	25 en la fase con destilación; la calibración previa admitió hasta 30-40
Tamaño de lote	256
Parada temprana	Paciencia de 5 épocas, restaurando los mejores pesos según la validación interna
Estandarización	StandardScaler ajustado únicamente con el subconjunto de entrenamiento de cada partición
PCA	Ajustado tras la estandarización, solo con entrenamiento, conservando el 95 % de la varianza explicada; aplicado luego a validación y prueba
Semillas finales	13, 17, 23, 29, 31
Particiones por semilla	3 particiones estratificadas
Conjunto de prueba	Reservado; sin uso en calibración ni selección de hiper-parámetros

Fase sin destilación: intervenciones sobre el MLP

Se comparó el MLP base y las variantes con reponderación (REW), PCA, PCA+REW, MinDiff, REW+MinDiff, PCA+MinDiff y adversarial debiasing; la reponderación se calculó según la ecuación (3) y los transformadores se ajustaron solo con los datos de entrenamiento de cada partición (Tabla 5). La intensidad de MinDiff se calibró sobre la rejilla {0.1, 0.5, 1.0, 2.0, 5.0} con un criterio de no inferioridad práctica respecto de la línea base del MLP; la Tabla 6 muestra que el mejor compromiso lo dio lambda igual a 5.0, con la menor FNR_DIF y desempeño no inferior.

Una configuración se consideró no inferior cuando su exactitud media y su recall medio en validación no cayeron más de 0.01 puntos absolutos respecto de la línea base del MLP, es decir, $ACC \geq ACC_ref - 0.01$ y $REC \geq REC_ref - 0.01$; este margen de no inferioridad se fijó antes de los experimentos. Una mejora estadísticamente significativa designa una diferencia con p de Holm inferior a 0.05 en la prueba pareada de Wilcoxon, y una mejora descriptiva designa una reducción del promedio o la mediana sin significancia tras la corrección.

Tabla 6. Selección de lambda de MinDiff en la fase sin destilación

Lambda MinDiff	FNR DIF	Accuracy	Recall	No inferior
5.0	0.000922	0.822073	0.998683	Sí
0.5	0.001373	0.821878	0.998324	Sí
2.0	0.001392	0.821780	0.997726	Sí
1.0	0.001677	0.821390	0.997247	Sí
0.1	0.002182	0.821585	0.995452	Sí

Modelos maestros, etiquetas suaves y ponderaciones de destilación

La destilación se apoyó en los dos maestros del estudio previo, que no se reentrenaron dentro del flujo neuronal: se cargaron como artefactos previos y se usaron para generar las probabilidades suaves de la clase positiva que alimentan al estudiante. La Tabla 7, reporta sus métricas operativas medidas dentro de este flujo experimental; estos valores soportan las ponderaciones de destilación y no reemplazan las cifras publicadas en el estudio original (sección 3.2), pues provienen de particiones y condiciones distintas.

Tabla 7. Métricas operativas de los maestros usadas para calcular los pesos de destilación

Maestro	Accuracy	Recall	FNR g0	FNR g1	FNR DIF	Interpretación
ABC+PHT	0.617831	0.899569	0.000000	0.000642	0.000542	Mayor paridad en FNR
RFC+REW+PCA	0.895162	0.980470	0.012255	0.021407	0.009152	Mayor exactitud operativa

Las ponderaciones de combinación no fueron aleatorias: se calcularon con reglas determinísticas a partir de la Tabla 7, con el promedio simple como control metodológico (Tabla 8). POND_ACC pondera por exactitud, POND_FAIR por el inverso de la disparidad y POND_MULTI combina ambos criterios.

Las cuatro ponderaciones son reglas determinísticas fijadas antes del experimento final y no producto de una optimización: el promedio simple actúa como control; las tres restantes se calculan en forma cerrada a partir de las métricas operativas de los maestros (proporcional a la exactitud; proporcional al inverso de la disparidad; y proporcional a la puntuación multiobjetivo exactitud/disparidad). Dichas métricas se calcularon sobre las predicciones de los maestros en el conjunto de entrenamiento (80 %) y se exportaron desde su entorno de contenedor de manera previa e independiente del flujo neuronal; el conjunto de prueba no participó en el cálculo de ningún peso, y los pesos permanecieron fijos durante todo el entrenamiento y la evaluación.

Tabla 8. Ponderaciones de combinación de maestros

Estrategia	w ABC+PHT	w RFC+REW+PCA	Regla de cálculo
PROM	0.500000	0.500000	Promedio simple: pesos fijos de control
POND_ACC	0.477149	0.522851	$w_i = Acc_i$ / suma(Acc)
POND_FAIR	0.944096	0.055904	$w_i = (1 / FNR_DIF)$ / suma(1 / FNR_DIF)
POND_MULTI	0.939067	0.060933	$score_i = Acc_i / FNR_DIF_i$; $w_i = score_i$ / suma(score)

Selección de hiper-parámetros de la fase con destilación

La fase con destilación tomó alpha mediante una rejilla de candidatos para controlar la proporción entre etiqueta real y señal destilada, y lambda se recalibró porque la intensidad óptima de MinDiff cambia cuando el estudiante aprende de etiquetas suaves. La Tabla 9 resume las rejillas y los valores obtenidos; los valores finales fueron alpha igual a 0.3, lambda de MinDiff igual a 0.1, lambda de adversarial debiasing igual a 0.5 y temperatura T igual a 3.0. El descenso de lambda de 5.0 a 0.1 entre fases indica que, bajo destilación, bastó una regularización más suave.

Tabla 9. Selección de hiper-parámetros en la fase con destilación

Parámetro	Valor	FNR DIF	Accuracy	Recall	No Inferior
alpha KD	0.3	0.000452	0.817685	0.999641	Sí
alpha KD	0.7	0.000753	0.821000	0.999402	Sí
alpha KD	0.5	0.000771	0.820415	0.998903	Sí
lambda MinDiff	0.1	0.000830	0.822853	0.998324	Sí
lambda MinDiff	5.0	0.000922	0.821780	0.999683	Sí
lambda MinDiff	2.0	0.001751	0.819928	0.997247	Sí
lambda MinDiff	1.0	0.001770	0.821088	0.997608	Sí
lambda MinDiff	0.5	0.002805	0.821488	0.999688	Sí
lambda adversarial	0.5	0.002916	0.821878	0.999931	Sí
lambda adversarial	0.1	0.003294	0.821878	0.997247	Sí
lambda adversarial	1.0	0.005061	0.821293	0.993058	Sí
lambda adversarial	2.0	0.014506	0.821293	0.982047	No

Evaluación final y análisis estadístico

El experimento final comparó ocho intervenciones sin destilación (MLP base, REW, PCA, PCA+REW, MinDiff, REW+MinDiff, PCA+MinDiff y adversarial debiasing) y doce con destilación, correspondientes a seis estrategias sin MinDiff y seis con MinDiff (maestros individuales ABC y RFC y las cuatro combinaciones de la Tabla 7). La fase sin destilación incluyó dos análisis de robustez: sensibilidad al umbral de decisión sobre una rejilla de 0.05 a 0.95, y robustez frente a ruido de etiquetas mediante inversión aleatoria de 0 % a 10 % de las etiquetas de entrenamiento con tres semillas.

Del estudio base se utilizaron los mismos conjuntos de datos correspondientes al 80 % entrenamiento 20 % prueba, y el conjunto de prueba (n = 4274) permaneció fijo. Por cada una de las cinco semillas (13, 17, 23, 29 y 31) se generaron tres particiones estratificadas de entrenamiento/validación dentro del 80 % de entrenamiento; la semilla controló tanto la partición como la inicialización de los pesos de la red. En cada una de las $5 \times 3 = 15$ ejecuciones por intervención, el modelo se entrenó con el subconjunto interno de entrenamiento, la parada temprana se controló con la validación interna y el modelo resultante se evaluó sobre el conjunto de prueba fijo.

Para cada intervención se conservaron las 15 ejecuciones obtenidas con las cinco semillas y las tres particiones. Estas ejecuciones se compararon de forma pareada con las del MLP base. Se calcularon intervalos BCa al 95 % para FNR_DIF, accuracy, recall, DPD y DI. Las diferencias frente al MLP base se analizaron mediante Wilcoxon pareado con corrección de Holm y se acompañaron del coeficiente rank-biserial. Finalmente, el plano FNR_DIF y accuracy se utilizó para identificar las soluciones no dominadas mediante la frontera de Pareto [43] [44] [46] [47].

Cada instrumento responde una pregunta de investigación: las pruebas pareadas de Wilcoxon con corrección de Holm evalúan si cada intervención difiere del MLP base (preguntas 1 y 3), y la frontera de Pareto identifica qué estrategia ofrece el mejor compromiso entre equidad y desempeño (pregunta 2).

Consideraciones éticas

Esta investigación corresponde a un análisis secundario de datos abiertos publicados por la Agencia para la Reincorporación y la Normalización (ARN) en el portal de Datos Abiertos del Gobierno de Colombia, utilizados conforme a los términos de la licencia de datos abiertos de dicho portal. La información se encuentra anonimizada en la fuente y no contiene datos de identificación personal; no hubo contacto con participantes ni recolección de información primaria. Por tratarse de un análisis secundario de datos públicos y anónimos, el estudio no requirió aprobación de comité de ética ni consentimiento informado.

Resultados

Los resultados siguen el orden del protocolo: primero la fase sin destilación y luego la fase con destilación. Los valores son promedios de las quince estimaciones en prueba, con intervalos de confianza BCa al 95 % para la métrica primaria.

Fase sin destilación: MLP, reponderación, PCA y MinDiff

La Tabla 10 muestra que, en la fase sin destilación, todas las variantes de preprocesamiento o MinDiff redujeron la FNR_DIF frente al MLP base (0.001821), con excepción de adversarial debiasing, que la incrementó (0.003541). La mejor equidad promedio la obtuvo MLP+PCA+MinDiff con lambda igual a 5.0 (0.000912, cerca de la mitad de la disparidad base), seguida de MLP+PCA+REW (0.000985); las variantes con reponderación quedaron en valores intermedios y con una exactitud semejante, y el MLP base retuvo la mayor exactitud de la fase (0.822976). MinDiff mejoró la equidad, especialmente junto con PCA, y ninguna intervención dominó simultáneamente todos los criterios.

Tabla 10. Resultados en prueba de las intervenciones sin destilación

Intervención	FNR DIF	IC95 FNR DIF	Accuracy	Recall	DPD	DI
MLP+PCA+MinDif (λ=5.0)	0.000912	0.000556-0.001220	0.821837	0.998181	0.007604	0.992168
MLP+PCA+REW	0.000985	0.000726-0.001151	0.821151	0.998200	0.007139	0.992836
MLP+REW+MinDif (λ=5.0)	0.001222	0.000958-0.001486	0.822555	0.998909	0.008181	0.991795
MLP+REW	0.001245	0.000860-0.001531	0.822617	0.998468	0.008317	0.991155
MLP+MinDif (λ=5.0)	0.001391	0.001081-0.001749	0.822524	0.998449	0.008214	0.991758
MLP+PCA	0.001556	0.001004-0.003074	0.821978	0.997741	0.009062	0.990907
MLP base	0.001821	0.001153-0.002944	0.822976	0.997836	0.010301	0.989667
MLP+AdvDebias (λ=0.1)	0.003541	0.002059-0.005861	0.822602	0.996439	0.013150	0.986615

La Figura 1 presenta el plano equidad-desempeño de esta fase, con FNR_DIF en el eje horizontal (menor es mejor) y exactitud en el vertical; los marcadores rellenos indican las soluciones no dominadas y la línea discontinua une la frontera de Pareto. Las soluciones no dominadas fueron MLP+PCA+MinDiff, MLP+REW+MinDiff, MLP+REW y el MLP base: cada una mejora la equidad de la siguiente a cambio de una fracción de exactitud, con un rango total de apenas 0.0011 entre los extremos de la frontera.

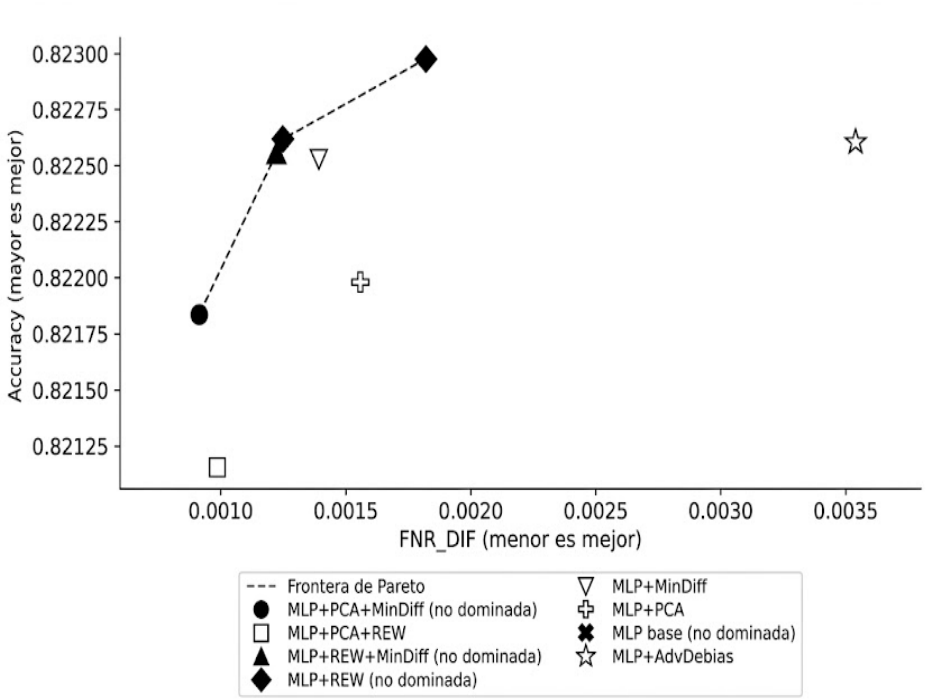

Figura 1. Plano FNR_DIF frente a accuracy de la fase sin destilación; los marcadores rellenos y la línea discontinua indican la frontera de Pareto

Los análisis de robustez se realizaron únicamente sobre las intervenciones de la fase sin destilación; en consecuencia, las conclusiones sobre la estabilidad del umbral y la tolerancia al ruido de etiquetas se circunscriben a esa fase y no se extrapolan al modelo destilado final. En la sensibilidad al umbral, la FNR_DIF de todas las intervenciones permaneció baja para umbrales menores o iguales a 0.5 y se degradó (aumentó) en un orden de magnitud a partir de 0.55 (el MLP base pasó de 0.0029 en 0.5 a 0.0183 en 0.55 y 0.0448 en 0.6), con REW+MinDiff y PCA+MinDiff conservando la menor disparidad; el umbral de 0.5 se ubica en una zona estable. Frente al ruido de etiquetas, con inversiones de hasta el 10 %, las variantes con MinDiff mantuvieron la FNR_DIF en torno a 0.001 con exactitud estable, mientras el MLP base mostró variabilidad (hasta 0.0037).

Fase con destilación: estrategias de maestros y MinDiff

Al construir el modelo mediante destilación, reportan mejores resultados de FNR_DIF como se muestra en la tabla 11. La combinación de la ponderación por exactitud con MinDiff (MLP_KD_POND_ACC_MD), obtuvo FNR_DIF de 0.000502 con exactitud de 0.819482 y recall de 0.999598, mientras que MLP_KD_PROM_MD alcanzó un valor similar (0.000503). El promedio simple sin MinDiff (MLP_KD_PROM, 0.000574) indica que la destilación por sí sola transfirió paridad aprendida por los maestros. Otra combinación, la destilación desde ABC en solitario fue la más débil (0.004287 sin MinDiff), pese a ser el maestro con mayor paridad, y las ponderaciones que concentraban el peso en él (POND_FAIR y POND_MULTI) quedaron rezagadas excepto cuando MinDiff participa en el entrenamiento.

Tabla 11. Comparación de estrategias de destilación

Intervención	MinDiff	FNR_DIF	IC95 FNR_DIF	Accuracy	Recall	DPD
MLP_KD_POND_ACC_MD	Sí	0.000502	0.000359-0.00574	0.819482	0.999598	0.005801
MLP_KD_PROM_MD	Sí	0.000503	0.000383-0.00623	0.818765	0.999579	0.004647
MLP_KD_PROM	No	0.000574	0.000383-0.000646	0.819077	0.999540	0.005525
MLP_KD_POND_FAIR_MD	Sí	0.000905	0.000408-0.001745	0.813867	0.996764	0.001346
MLP_KD_POND_ACC	No	0.000999	0.000407-0.003072	0.817501	0.999368	0.004077
MLP_KD_RFC_MD	Sí	0.001005	0.000743-0.001292	0.822430	0.998985	0.008615
MLP_KD_RFC	No	0.001029	0.000743-0.001315	0.822852	0.998880	0.008499
MLP_KD_POND_MULTI_MD	Sí	0.001483	0.000690-0.002683	0.813695	0.996669	0.001634
MLP_KD_POND_FAIR	No	0.002429	0.001000-0.005101	0.812479	0.993548	0.003699
MLP_KD_POND_MULTI	No	0.002827	0.001394-0.005821	0.811683	0.994026	0.003150
MLP_KD_ABC_MD	Sí	0.003002	0.000666-0.009615	0.812135	0.994582	0.003386
MLP_KD_ABC	No	0.004287	0.001174-0.010139	0.808158	0.985392	0.005631

La Figura 2 compara la FNR_DIF de las seis estrategias con y sin MinDiff, con la línea discontinua en el valor del MLP base de esta fase (0.000982). MinDiff mejoró o mantuvo la equidad en todas las estrategias, con mayor ganancia donde más se necesitaba (ABC, POND_FAIR y POND_MULTI), y las tres mejores configuraciones quedaron por debajo del MLP base.

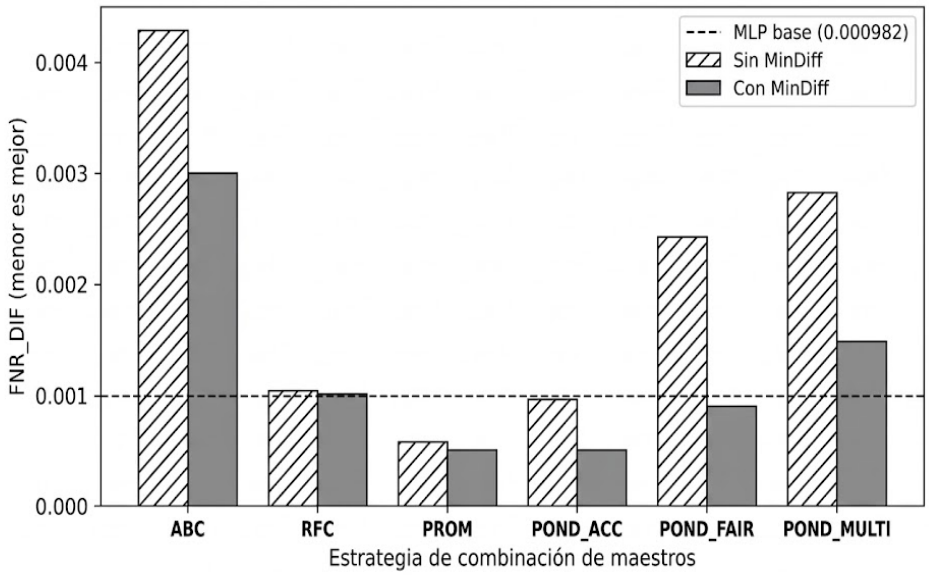

Figura 2. FNR_DIF por estrategia de combinación de maestros, con y sin MinDiff

La Figura 3, presenta la frontera de Pareto de la fase con destilación. Solo tres soluciones resultaron no dominadas: MLP_KD_POND_ACC_MD, orientada a equidad (FNR_DIF de 0.000502, exactitud de 0.819482 y recall de 0.999598); el MLP base, con la mayor exactitud entre los modelos de disparidad baja (0.000982, 0.822914 y 0.998736); y MLP+AdvDebias con lambda igual a 0.5, que maximiza la exactitud a costa de casi triplicar la disparidad (0.002871, 0.823117 y 0.996669). Para un decisor cuyo criterio dominante sea la no discriminación en falsos negativos, la configuración destilada con MinDiff es la elección: cuesta 0.0034 puntos de exactitud (de 0.822914 a 0.819482) frente al MLP base a cambio de reducir la disparidad a cerca de la mitad y elevar el recall.

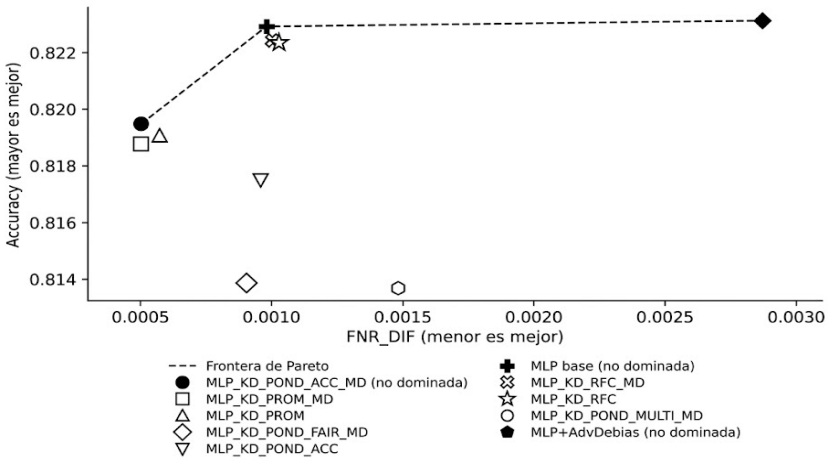

Figura 3. Plano FNR_DIF frente a accuracy de la fase con destilación; los marcadores rellenos y la línea discontinua indican la frontera de Pareto

Métricas por grupo y verificación estadística

La Tabla 12 y la Figura 4 desagregan la FNR por grupo para los modelos de la fase con destilación; la FNR_DIF corresponde al promedio por ejecución. Las configuraciones destiladas llevan la FNR del grupo de mujeres (g0) a cero o casi cero y reducen la del grupo de hombres (g1) a valores entre 0.000502 y 0.000574, alrededor de un tercio del valor del MLP base para ese grupo (0.001363); las tasas de selección también se acercan, con la DPD descendiendo de 0.008751 a valores entre 0.004647 y 0.005801 y el DI acercándose a uno.

Tabla 12. Métricas por grupo para los modelos clave de la fase con destilación

Intervención	FNR g0	FNR g1	FNR_DIF	SR g0	SR g1	DPD	DI
MLP_KD_POND_ACC_MD	0.000000	0.000502	0.000502	9.999309	0.993508	0.005801	0.994195
MLP_KD_PROM_MD	0.000096	0.000502	0.000503	0.999050	0.994403	0.004647	0.995348
MLP_KD_PROM	0.000000	0.000574	0.000574	0.999396	0.993870	0.005525	0.994471
MLP base	0.000863	0.001363	0.000482	0.996991	0.988140	0.008751	0.991223
MLP+AdvDebias (λ=0.5)	0.001343	0.003827	0.002871	0.996459	0.983876	0.012584	0.987373

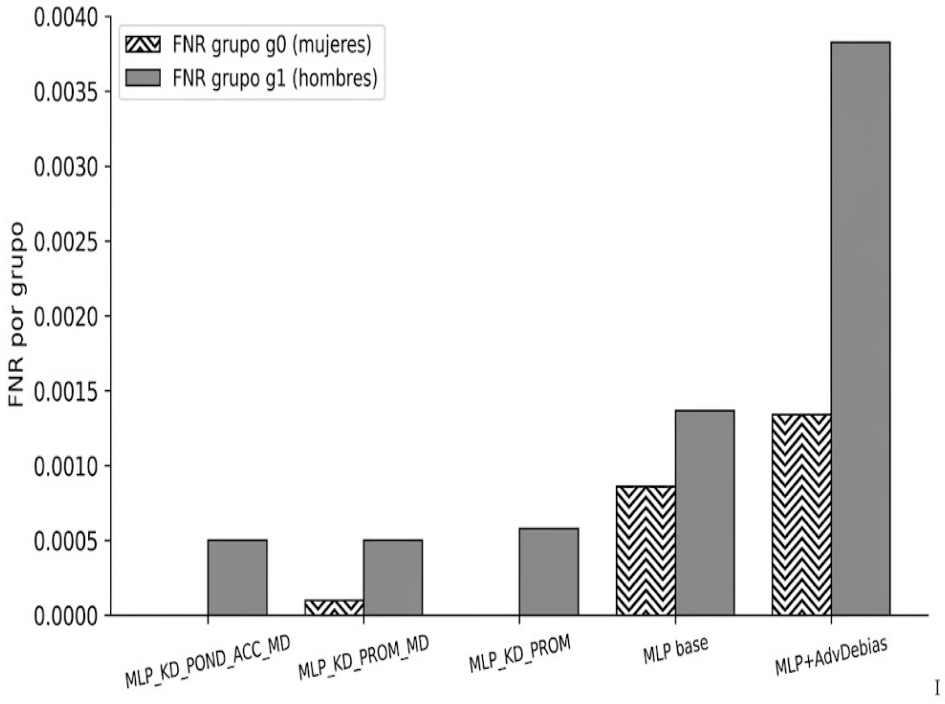

Figura 4. FNR por grupo sensible en los modelos clave de la fase con destilación

La Tabla 13 presenta las pruebas pareadas de la mejor configuración destilada frente al MLP base. La reducción de FNR_DIF tuvo una diferencia mediana de -0.000362 con un tamaño de efecto medio y consistente entre las quince comparaciones (rank-biserial de -0.60; p = 0.0075; p de Holm = 0.1432). Tras la corrección de Holm resultaron estadísticamente significativas la reducción de DPD (p = 0.0068, rank-biserial de -0.73), el aumento de recall (p = 0.0101, rank-biserial de 1.00) y la reducción de exactitud (p = 0.0123, rank-biserial de -1.00), que cuantifica el costo del intercambio: la ganancia en igualdad de asignación y en sensibilidad es sistemática, con una pérdida moderada y acotada de exactitud.

Tabla 13. Pruebas pareadas de MLP_KD_POND_ACC_MD frente al MLP base

Métrica	Diferencia mediana	Rank-biserial	p	p Holm	Significativa (0.05)
FNR DIF	−0.000362	−0.600000	0.007537	0.143212	No
DPD	−0.002549	−0.733333	0.000427	0.005836	Sí
Accuracy	−0.002574	−1.000000	0.000847	0.012288	Sí
Recall	0.000574	1.000000	0.000560	0.010073	Sí

Comparación con el estudio base

Frente al estudio base [15], el estudiante destilado conservó y afinó la paridad del mejor modelo preprocesado. La combinación de MLP_KD_POND_ACC_MD alcanzó una FNR_DIF de 0.000502, inferior a la del maestro ABC+PHT medida en las mismas condiciones (0.000542) y del mismo orden que la reportada en el estudio original para ese modelo (0.0004), con exactitud superior a la del maestro (0.819482 frente a 0.817831) y sensibilidad equivalente (0.999598 frente a 0.999569). Frente al maestro que combinó RFC+REW+PCA, se redujo la discriminación en más de un orden de magnitud (0.000502 frente a 0.009152), con una disminución de 0.0767 puntos de exactitud (de 0.896162 del maestro a 0.819482 del estudiante). La destilación igualó y mejoró levemente, en un modelo neuronal, el equilibrio paridad-exactitud que el estudio base había alcanzado con clasificadores tradicionales. Estos resultados permitieron afirmar que mezclar señales por promedio simple o ponderación por exactitud; y MinDiff aportó una mejora adicional de bajo costo con baja intensidad sobre estudiantes destilados, mientras que su ganancia aislada, sin destilación, fue menor.

Discusión

El preprocesamiento estabiliza y la destilación transfiere

Los resultados se alinean con la literatura en cuanto a que el preprocesamiento aumenta la equidad con pérdidas pequeñas de desempeño, mientras que las intervenciones durante el entrenamiento permiten establecer explícitamente el criterio de igualdad deseado [2] [4] [3]. En la fase sin destilación, la reponderación y PCA redujeron la disparidad del MLP con costos pequeños de rendimiento, mientras que MinDiff con PCA obtuvo la mejor equidad promedio [5] [34] [35]. Esto respalda la afirmación que el problema está en los datos y que los algoritmos aprenden de ellos, por lo tanto, primero se logra tener modelos más equitativos, cuando las intervenciones tienen en cuenta los datos [15] [48] [27] y si las intervenciones de procesamiento se combinan con el procesamiento incrementa la equidad [24] [30] [32] [33].

Uno de los principales hallazgos, proviene de la intervención con destilación: la destilación desde maestros mitigados redujo la FNR_DIF frente al MLP base y conservando un recall alto, incluso sin MinDiff. Esto se ajusta con el proceso documentado en donde se entrenan o selecciona maestros justos para que el estudiante herede representaciones equitativas [19] y con la ventaja que la destilación ofrece una optimización más estable que la optimización directa de la equidad [39]. Al mismo tiempo, matiza esa literatura: la mejor configuración no fue concentrar el peso en el maestro más equitativo, sino combinar señales por promedio o por exactitud. La destilación desde ABC produjo los estudiantes más débiles y variables, lo que indica que una señal casi saturada en recall aporta poca información de frontera de decisión; un maestro con mayor exactitud aporta información si su influencia se controla, para que se ajusten con los marcos de múltiples maestros con roles diferenciados [40] [49] y con la precaución de que el estudiante hereda solo parcialmente las propiedades del maestro cuando existe brecha entre ambos [18].

MinDiff depende de la señal de aprendizaje

Si bien, como ya se mencionó las intervenciones en procesamiento permiten tener mayor control sobre la compensación equidad – rendimiento, el comportamiento de MinDiff, al ser una intervención durante el entrenamiento del modelo, permitió establecer que su efectividad depende de la señal con la que aprende el modelo: la intensidad óptima pasó de 5.0 sin destilación a 0.1 con destilación, cuando los estudiantes ya reciben etiquetas suaves de maestros mitigados. Este comportamiento sugiere, que cuando el criterio de equidad ya viene codificado en la señal destilada, la regularización explícita necesita menor intensidad, este resultado se alinea con los estudios que condicionan la utilidad de las intervenciones in-processing a la capacidad del modelo y a la forma de la regularización [23] [35] [34].

El contraste con adversarial debiasing, cuyas intervenciones penalizan la predictibilidad del atributo sensible desde las representaciones, y no la diferencia de errores entre grupos, se pudo evidenciar que fue la única variante que empeoró la disparidad de falsos negativos en ambas fases, aunque con lambda igual a 0.5 entró a la frontera de Pareto por su exactitud. Esta limitación se debe a que esta técnica no permite establecer la métrica del dominio a mejorar, para que las intervenciones estén alineadas con el criterio de equidad que la decisión de la política pública desee garantizar [2] [13].

¿Preprocesamiento o procesamiento?

Desde el punto de vista de este artículo que consiste en evaluar como la combinación de intervenciones puede mejorar la equidad se podría establecer que los resultados sugieren que no existe una superioridad de una familia de intervenciones sobre la otra, pero sí que su combinación de forma secuencial muestra mejores resultados. El mejor resultado global surgió de combinar ambas familias: el preprocesamiento construyó maestros equitativos en el estudio base, la destilación transfirió esa equidad al estudiante neuronal y MinDiff la afinó (0.000502). Lo que permite afirmar que ambos enfoques son complementarios. Pero al querer compararlas por separado sobre la misma arquitectura, el preprocesamiento (reponderación y PCA) y la regularización in-processing (MinDiff) alcanzaron niveles de equidad similares, con el mejor valor individual en su combinación (PCA+MinDiff, 0.000912).

Lectura estadística de los resultados

Las estrategias de destilación con maestros mitigados alcanzan mejoras prácticas en equidad con desempeño semejante; la magnitud de la mejora, su estabilidad y su costo en exactitud quedan cuantificados, lo que permite seleccionar la configuración según la prioridad del tomador de decisiones [46] [45]. Esto es respaldado por la mejora del promedio y la mediana de FNR_DIF frente al MLP base con un tamaño de efecto medio (mejora descriptiva, no significativa tras Holm: p ajustado = 0.1432), y logra mejoras estadísticamente significativas tras la corrección en DPD y recall, con un costo significativo, moderado y cuantificado en exactitud [43] [42] [45].

Implicaciones para la decisión de política

Estos resultados ayudan a la toma de decisiones porque si el criterio es minimizar la disparidad de falsos negativos, la elección es la configuración destilada con ponderación por exactitud y MinDiff de baja intensidad: cuesta 0.0034 puntos de exactitud frente al MLP base con ganancia en recall, lo que implica reconocer a más personas elegibles de ambos sexos y repartir los errores restantes de forma equilibrada. Si el criterio dominante es la exactitud con disparidad contenida, la alternativa es el MLP base o las variantes con reponderación sin destilación [3] [28].

Limitaciones

Los resultados dependen de la distribución del conjunto de datos utilizado, específicamente en problema en donde exista predominio de la clase positiva y del grupo masculino, por lo que su transferencia a otros casos de uso requiere realizar su propia validación. Adicionalmente, las etiquetas provienen de decisiones históricas, de modo que la equidad medida es relativa a esas etiquetas y no elimina posibles sesgos de medición previos [7] [27].

A lo anterior se suman cuatro limitaciones adicionales. Primera, el estudio emplea un único conjunto de datos administrativo colombiano y carece de validación externa en otros programas, poblaciones o países. Segunda, los datos reflejan un periodo específico del proceso de reintegración, por lo que cambios en la distribución a lo largo del tiempo (deriva temporal) podrían alterar tanto el desempeño como los patrones de disparidad aquí observados. Tercera, la comparación se restringió a una única arquitectura neuronal fija (MLP); no se evaluaron otras arquitecturas ni enfoques de aprendizaje profundo alternativos, decisión que aísla el efecto de las intervenciones, pero limita la generalidad arquitectónica de los hallazgos. Cuarta, aunque el sexo se excluyó de los predictores, no se controló la posible presencia de variables proxy correlacionadas con él.

Existen otras limitaciones que invitan a realizar trabajos futuros, por ejemplo: que el caso de uso abordado fue una tarea de clasificación y atributo protegido binario; quedando pendiente abordarlo, pero con interseccionalidad [39]. Otro experimento como trabajo futuro consiste en que las estrategias de combinación de maestros fueron determinísticas y estáticas; las señales dependientes de instancia pueden ser también explorados para determinar su posible aporte [49] [32] [38]. La sensibilidad al umbral se verificó en la fase sin destilación y mostró una región estable alrededor de 0.5; quedaría pendiente su réplica en la fase con destilación y las intervenciones de posprocesamiento por grupo [35].

Conclusiones

Esta investigación produjo una contribución metodológica que consistió en aprovechar el aprendizaje adquirido por un modelo tradicional para utilizarlo como maestro equilibrado; de forma que puedan generar etiquetas suaves sin reentrenarlos ni usar el conjunto de prueba; generando las combinaciones mediante cálculos de ponderaciones; permitiendo entrenar estudiantes con destilación, con y sin regularización de equidad; y establecer el conjunto de combinaciones que mejor mantengan la compensación entre rendimiento y equidad en la salida del modelo, aislando la contribución de cada técnica, proporcionando un procedimiento para evaluar criterios de no discriminación bajo las condiciones estudiadas.

La mejor intervención combino señales complementarias de paridad y exactitud con reglas de ponderación. Los resultados sugieren que la equidad alcanzada por clasificadores tradicionales mediante preprocesamiento puede transferirse a una red neuronal mediante destilación de conocimiento: los estudiantes destilados redujeron las proporciones de falsos negativos entre sexos, con sensibilidad cercana a uno, y MinDiff con baja intensidad redujo esa paridad con un costo moderado y cuantificado en exactitud.

Los resultados muestran, bajo las condiciones evaluadas y sin validación externa, que las intervenciones propuestas en la literatura, validadas con conjuntos de datos populares y de uso frecuente entre los investigadores, pueden aplicarse también en contextos reales, como el caso colombiano de asignación de beneficios sociales. Pero no todas las intervenciones resultan efectivas, por lo que es necesario seleccionar aquellas que mejor se alineen con el criterio de equidad que se pretenda garantizar.

Bajo las condiciones evaluadas un único conjunto de datos administrativo colombiano, una arquitectura fija y sin validación externa los resultados muestran una reducción descriptiva y consistente de la disparidad de falsos negativos con la destilación desde maestros mitigados (de 0.000982 a 0.000502 en promedio, no significativa tras la corrección de Holm), junto con mejoras estadísticamente significativas en la paridad demográfica y la sensibilidad, y un costo significativo, moderado y cuantificado en exactitud. Sobre esa evidencia se sustenta la recomendación práctica de una estrategia escalonada preprocesamiento en los maestros, transferencia por destilación y regularización MinDiff de baja intensidad cuya generalización a otros dominios requiere validación empírica adicional.

Declaraciones finales

Contribución de los autores (Taxonomía CRediT)

Alveiro Alonso Rosado Gomez: conceptualización, metodología, investigación y redacción del borrador original, análisis formal, curación de datos, visualización y revisión y edición.

Financiación

Esta investigación no recibió financiación específica de entidades públicas, comerciales o sin ánimo de lucro.

Conflictos de intereses

El autor declara que no existen conflictos de intereses relacionados con este trabajo.

Disponibilidad de datos

Los datos utilizados provienen del conjunto "Estadísticas de las personas desmovilizadas que han ingresado al proceso de reintegración", publicado por la Agencia para la Reincorporación y la Normalización (ARN) en el portal de Datos Abiertos del Gobierno de Colombia bajo licencia pública, disponible en: https://www.datos.gov.co/Inclusi-n-Social-y-Reconciliaci-n/ESTAD-STICAS-DE-LAS-PERSONAS-DESMOVILIZADAS-QUE-HA/39pj-dba6/about_data. El código y los artefactos experimentales están disponibles a solicitud al autor de correspondencia.

Consideraciones éticas

Esta investigación corresponde a un análisis secundario de datos abiertos, anonimizados y de acceso público, publicados por la ARN en el portal de Datos Abiertos del Gobierno de Colombia, sin contacto con participantes ni recolección de información primaria, por lo que no requirió aprobación de comité de ética ni consentimiento informado.

Uso de inteligencia artificial generativa

Se utilizó la herramienta de inteligencia artificial generativa Claude (Anthropic) exclusivamente como apoyo en la adecuación del manuscrito a la plantilla editorial y en la revisión de estilo y redacción en español. La herramienta no participó en el diseño metodológico, la ejecución de los experimentos, el análisis o la interpretación de los resultados, ni en la generación de contenido científico. Todo el contenido es de autoría exclusiva del investigador, quien revisó y validó la totalidad del texto y asume plena responsabilidad por él.

Referencias

[1]T. Hellstrom, V. Dignum y S. Bensch, «Bias in Machine Learning What is it Good for?,» arXiv, pp. 1-8, 2020. doi:https://doi.org/10.48550/arXiv.2004.00686

[2]K. Makhlouf, S. Zhioua y C. Palamidessi, «Machine learning fairness notions: Bridging the gap with real-world applications,» Information Processing & Management, pp. 1-32, 2021. doi:https://doi.org/10.1016/j.ipm.2021.102642

[3]S. Caton y C. Haas, «Fairness in Machine Learning: a Survey,» arXiv, pp. 1-33, 2020. doi:https://doi.org/10.48550/arXiv.2010.04053

[4]N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman y A. Galstyan, «A Survey on Bias and Fairness in Machine Learning,» arXiv, pp. 1-31, 2019. doi:https://doi.org/10.48550/arXiv.1908.09635

[5]M. Hort, Z. Chen, J. Zhang, M. Harman y F. Sarro, «Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey,» ACM Journal on Responsible Computing, pp. 1-52, 2024. doi: https://doi.org/10.1145/3631326

[6] A. Agarwal y H. Agarwal, «A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle,» AI and Ethics, pp. 299-314, 2024. doi:https://doi.org/10.1007/s43681-023-00266-9

[7] A. Olteanu, C. Castillo, F. Diaz y E. Kıcıman, «Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries,» Frontiers in Big Data, pp. 1-33, 2019. doi:https://doi.org/10.3389/fdata.2019.00013

[8] S. Verma y J. Rubin, «Fairness Definitions Explained,» de International Conference on Software Engineering, Gotemburgo, 2018 doi:https://doi.org/10.1145/3194770.3194776.

[9] R. Burke, «Multisided Fairness for Recommendation,» de Fairness, Accountability, and Transparency in Machine Learning, Halifax, 2017. doi:https://doi.org/10.48550/arXiv.1707.00093

[10] S. Barocas, E. Bradley, V. Honavar y F. Provost, «Big Data, Data Science, and Civil Rights,» Computing Community Consortium (CCC), pp. 1-8, 2017. doi:https://doi.org/10.48550/arXiv.1706.03102

[11] I. Zliobaite, «Fairness-aware machine learning: a perspective,» ArXiv, pp. 1-10, 2017. doi:https://doi.org/10.48550/arXiv.1708.00754

[12]B. Richardson y J. Gilbert, «A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions,» arXiv, pp. 1-28, 2021. doi:https://doi.org/10.48550/arXiv.2112.05700

[13]D. Danks y A. London, «Algorithmic Bias in Autonomous Systems,» de Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, 2017. doi:https://doi.org/10.24963/ijcai.2017/654

[14]T. Salazar, M. Seoane, H. Araújo y P. Henriques, «FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes,» IEEE Access, pp. 81370-81379, 2021. doi:https://doi.org/10.1109/ACCESS.2021.3084121

[15]A. Rosado, M. L. Calderón y O. Espinosa, «Data preprocessing to improve fairness in machine learning models: An application to the reintegration process of demobilized members of armed groups in Colombia,» Applied Soft Computing, pp. 1-14, 2024. doi:https://doi.org/10.1016/j.asoc.2023.111193

[16]G. Hinton, O. Vinyals y J. Dean, «Distilling the Knowledge in a Neural Network,» arXiv, pp. 1-9, 2015. doi:https://doi.org/10.48550/arXiv.1503.02531

[17]J. Gou, B. Yu, S. Maybank y D. Tao, «Knowledge Distillation: A Survey,» International Journal of Computer Vision, p. 1789–1819, 2021. doi:https://doi.org/10.1007/s11263-021-01453-z

[18]A. Mohammadshahi y Y. Ioannou, «What is Left After Distillation? How Knowledge Transfer Impacts Fairness and Bias,» arXiv, pp. 1-22, 2024. doi:https://doi.org/10.48550/arXiv.2410.08407

[19]H. Tian, B. Liu, T. Zhu, W. Zhou y P. S. Yu, «Distilling Fair Representations From Fair Teachers,» IEEE Transactions on Big Data, pp. 1419-1433, 2025.

[20]P. Lahoti, K. Gummadi y G. Weikum, «Operationalizing individual fairness with pairwise fair representations,» Proceedings of the VLDB Endowment, pp. 506-518, 2019. doi:https://doi.org/10.48550/arXiv.1907.01439

[21]J. Atwood, T. Tian, B. Packer, M. Deodhar, J. Chen, A. Beutel, F. Prost y A. Beirami, «Towards A Scalable Solution for Improving Multi-Group Fairness in Compositional Classification,» arXiv, pp. 1-10, 2023. doi: https://doi.org/10.48550/arXiv.2307.05728

[22]M. Wan, D. Zha, N. Liu y N. Zou, «In-Processing modeling techniques for machine learning fairness: A survey,» ACM Trans. Knowl. Discov. Data, pp. 1-27, 2023. doi:https://doi.org/10.1145/3551390

[23]A. K. Veldanda, I. Brugere, J. Chen, S. Dutta, A. Mishler y S. Garg, «Fairness via In-Processing in the Over-parameterized Regime: A Cautionary Tale,» arXiv, pp. 1-14, 2022. doi:https://doi.org/10.48550/arXiv.2206.14853

[24]Z. Chen, J. Zhang, F. Sarro y M. Harman, «A comprehensive empirical study of bias mitigation methods for machine learning classifiers,» arXiv, pp. 1-30, 2023. doi:https://doi.org/10.48550/arXiv.2207.03277

[25]X. He, K. Zhao y X. Chu, «AutoML: A survey of the state-of-the-art,» Knowledge-Based Systems, pp. 1-36, 2020. doi:https://doi.org/10.1016/j.knosys.2020.106622

[26]M. Zhou, V. Abhishek, T. Derdenger, J. Kim y K. Srinivasan, «Bias in Generative AI,» arXiv, pp. 1-21, 2024. doi:https://doi.org/10.48550/arXiv.2403.02726

[27]H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," ACM Conf. Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1-9, 2021, doi: 10.1145/3465416.3483305

[28]A. Nielsen, Practical Fairness Achieving Fair and Secure Data Models, Sebastopol: O'Reilly, 2020.

[29]F. Kamiran y T. Calders, «Data preprocessing techniques for classification without discrimination,» Knowledge and Information Systems, vol. 33, p. 1–33, 2012. doi:https://doi.org/10.1007/s10115-011-0463-8

[30]M. Kamani, F. Haddadpour, R. Forsati y M. Mahdavi, «Efficient fair principal component analysis,» Machine Learning, pp. 3671-3702, 2022. doi:https://doi.org/10.1007/s10994-021-06100-9

[31]S. Masís, Interpretable machine learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples, Birmingham: Packt, 2021.

[32]A. Tawakuli y T. Engel, «Make Your Data Fair: A Survey of Data Preprocessing Techniques that Address Biases in Data Towards Fair AI,» Journal of Engineering Research, pp. 2307-1877, 2025. doi:https://doi.org/10.1016/j.jer.2024.06.016

[33]X. Wang, Y. Zhang y R. Zhu, «A Brief Review on Algorithmic Fairness,» Management System Engineering, 2022.

[34]J. Atwood, N. Scherrer, P. Lahoti, A. Balashankar, F. Prost y A. Beirami, «Inducing Group Fairness in Prompt-Based Language Model Decisions,» arXiv, pp 1-11, 2024. doi:https://doi.org/10.48550/arXiv.2406.16738

[35]A. Tifrea, P. Lahoti, B. Packer, Y. Halpern, A. Beirami y F. Prost, «FRAPPE: A Group Fairness Framework for Post-Processing Everything,» arXiv, pp. 1-23, 2023. doi:https://doi.org/10.48550/arXiv.2312.02592

[36]P. Dantas, W. Sabino da Silva, L. Cordeiro y C. Carvalho, «Distilling Fair Representations From Fair Teachers,» Applied Intelligence, pp. 11804–11844, 2024. doi:https://doi.org/10.1109/TBDATA.2024.3460532

[37]S. Wu, X. Luo, J. Liu y Y. Deng, «Knowledge distillation with adapted weight,» Statistics, pp. 470-497, 2025. doi:https://doi.org/10.1080/02331888.2025.2451944

[38]S. Zhao, R. Duan, X. Wang y W. Xingxing, «Improving adversarial robust fairness via anti-bias soft label distillation,» de NeurIPS, Vancouver, 2024.

[39]F. Sikder, R. Ramachandranpillai, D. de Leng y F. Heintz, «Promoting intersectional fairness through knowledge distillation,» IOS Press, pp. 3431-3434, 2025. doi:https://doi.org/10.3233/FAIA251214

[40]M. Masroor, T. Hassan, Y. Tian, K. Wells, D. Rosewarne, T.-T. Do y G. Carneiro, «Fair Distillation: Teaching Fairness from Biased Teachers in Medical Imaging,» arXiv, pp. 1-16, 2024. doi:https://doi.org/10.48550/arXiv.2411.11939

[41]X. Yue, N. Mou, Q. Wang y L. Zhao, «Revisiting adversarial robustness distillation from the perspective of robust fairness,» de International Conference on Neural Information Processing Systems, New Orleans, 2023. https://doi.org/10.52202/075280-1323

[42]S. a. H. F. García, «An Extension on Statistical Comparisons of Classifiers over Multiple Data Sets for All Pairwise Comparisons,» Journal of Machine Learning Research, pp. 2677-2694, 2008.

[43]J. Demšar, «Statistical Comparisons of Classifiers over Multiple Data Sets,» Journal of Machine Learning Research, pp. 1-30, 2006.

[44]S. Holm, «A Simple Sequentially Rejective Multiple Test Procedure,» Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979.

[45]R. Wasserstein y N. Lazar, «The ASA Statement on p-Values: Context, Process, and Purpose,» The American Statistician, pp. 129–133, 2016. doi:https://doi.org/10.1080/00031305.2016.1154108

[46]D. S. Kerby, «The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation,» Comprehensive Psychology, pp. 1-9, 2014. doi:https://doi.org/10.2466/11.IT.3.1

[47]F. Wilcoxon, «Individual Comparisons by Ranking Methods,» de Breakthroughs in Statistics: Methodology and Distribution, New York, Springer, 1992, pp. 196-202.

[48]R. Schwartz, L. Down, A. Jonas y E. Tabassi, A Proposal for Identifying and Managing Bias in Artificial Intelligence, Gaithersburg: National Institute of Standards and Technology, 2021. doi:https://doi.org/10.6028/NIST.SP.1270-draft

[49]J. Li, X. Feng, T. Gu y L. Chang, «Dual-Teacher De-Biasing Distillation Framework for Multi-Domain Fake News Detection,» de International Conference on Data Engineering (ICDE), Utrecht, 2024. https://doi.org/10.1109/ICDE60146.2024.00279