Auditoría Secuencial de Equidad con Acceso Limitado al Modelo

Ioannis Pitsiorlas[1*][0000−0003−0749−013X], Martha V. Sourla[2*][0009−0003−9113−7490], y Marios Kountouris[2][0000−0003−1143−080X]

[1] Departamento de Sistemas de Comunicación, EURECOM, Francia
[2] Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI), Universidad de Granada, España
[*]Estos autores contribuyeron por igual.

Resumen Las evaluaciones externas desempeñan un papel cada vez más central en la gobernanza de los sistemas de IA; sin embargo, los auditores independientes suelen tener acceso limitado a los modelos desplegados y deben recurrir a interacciones basadas en consultas. La mayoría de los métodos de evaluación de equidad suponen conjuntos de datos estáticos y pruebas con tamaño muestral fijo, poco adecuadas para acumular evidencia secuencial bajo presupuestos de consulta limitados. En este trabajo, formulamos la auditoría de equidad como un problema de prueba de hipótesis secuencial con tolerancia explícita y acceso limitado a las salidas del modelo. Desarrollamos un marco basado en la razón de verosimilitudes generalizada y reglas de parada explícitas, aplicable a auditorías de paridad estadística e igualdad de oportunidades basadas en decisiones discretas, y extensible a auditorías proxy con puntuaciones y logits cuando se dispone de salidas más informativas. Los resultados muestran que tanto la métrica de equidad como el nivel de acceso afectan sustancialmente a la eficiencia: las salidas más ricas pueden reducir el coste de las consultas en algunas métricas y regímenes operativos, aunque ofrecen ganancias limitadas cerca del umbral. El marco propuesto permite realizar auditorías secuenciales de equidad bajo restricciones de despliegue realistas.

Keywords: Prueba de hipótesis secuencial · Equidad algorítmica · Auditoría de equidad · Acceso limitado a las salidas del modelo

1. Introducción

Los sistemas de decisión algorítmica están cada vez más sujetos a supervisión normativa. Marcos como la Ley de IA de la UE [5] y la Ley Local 144 de Nueva York [10] refuerzan las obligaciones de gestión de riesgos, transparencia y auditoría de los sistemas de IA desplegados. Sin embargo, los auditores externos rara vez acceden a los parámetros del modelo, los datos de entrenamiento o la documentación interna. La auditoría suele realizarse mediante consultas al sistema, observando sus salidas para entradas seleccionadas y extrayendo conclusiones a partir de evidencia limitada [3]. En este artículo, distinguimos entre auditorías exactas de caja negra de restricciones de equidad basadas en decisiones y auditorías *proxy*, sensibles al régimen de acceso, que emplean salidas más informativas,

como puntuaciones (*scores*) o logits, para estudiar su efecto en la acumulación secuencial de evidencia.

La mayoría de los métodos de evaluación de equidad se basan en métricas estáticas, como la Paridad Estadística (*Statistical Parity, SP*) o la Igualdad de Oportunidades (*Equal Opportunity, EO*) [6, 7], y en pruebas de hipótesis con tamaño muestral fijo, diseñadas para conjuntos de datos previamente recopilados. Aunque una línea de trabajo creciente estudia la auditoría con acceso limitado [3], la prueba secuencial de hipótesis ofrece herramientas especialmente adecuadas para estos escenarios: los métodos iniciados por Wald [11] permiten detener el procedimiento de forma temprana controlando las tasas de error, y desarrollos recientes en inferencia uniforme en el tiempo extienden estas ideas a entornos adaptativos [8]. Asimismo, trabajos recientes formulan la auditoría como un problema de decisión estadística sobre salidas observables [3]. No obstante, los enfoques existentes no proporcionan procedimientos secuenciales concretos con reglas de parada explícitas ni, hasta donde sabemos, un marco basado en la razón de verosimilitudes generalizada (*Generalized Likelihood Ratio, GLR*) para auditorías de equidad con tolerancia explícita y acceso limitado.

Formalizamos la auditoría de equidad como un problema de prueba de hipótesis secuencial. Dada una métrica de equidad $g(f)$ y una tolerancia normativa δ, consideramos la prueba compuesta

$$H_0 : g(f) \leq \delta \qquad \text{frente a} \qquad H_1 : g(f) > \delta.$$

A diferencia de los procedimientos con tamaño muestral fijo, el protocolo acumula evidencia secuencialmente y finaliza cuando esta resulta suficiente o se agota un presupuesto predefinido de consultas. La formulación refleja escenarios normativos en los que pequeñas disparidades pueden ser admisibles y el cumplimiento se define respecto a un umbral de tolerancia, no a una disparidad nula.

El estudio empírico revela tres fenómenos principales. Primero, la tolerancia explícita genera fuertes efectos de frontera: los modelos con disparidades cercanas al umbral suelen producir resultados inconcluyentes bajo presupuestos realistas. Segundo, el acceso a salidas más informativas puede reducir sustancialmente el coste de la auditoría para SP y ciertos regímenes operativos, aunque las ganancias no son uniformes entre criterios de equidad. Tercero, las auditorías de EO suelen requerir más muestras que las de SP porque operan sobre el conjunto condicionado a $Y = 1$, generalmente más pequeño.

2. Metodología

Estudiamos la auditoría de equidad como un problema de inferencia estadística secuencial derivado de la interacción con un clasificador desplegado a través de consultas sobre un conjunto de auditoría fijo y reservado. El objetivo es determinar si el modelo satisface una restricción de equidad con tolerancia explícita, minimizando el número de consultas requeridas para alcanzar una decisión fiable.

2.1. Objetivo de la Auditoría y Definiciones de Equidad

Sea $f : \mathcal{X} \to \{0, 1\}$ un clasificador binario desplegado, con cada entrada $x \in \mathcal{X}$ asociada a un atributo sensible $A \in \{a, b\}$. Sea Q la distribución de

referencia sobre (X, A, Y), con $Y \in \{0, 1\}$ la etiqueta real disponible en el conjunto etiquetado. Un modelo se declara conforme con respecto a una tolerancia normativa $\delta > 0$ si $g_Q(f) \leq \delta$. La auditoría basada únicamente en decisiones evalúa directamente SP y EO basadas en decisiones discretas; cuando se dispone de salidas más ricas, se consideran auditorías *proxy* auxiliares que preservan la misma estructura de condicionamiento pero operan sobre puntuaciones o logits.

Paridad estadística (SP) [6] Definimos las tasas positivas condicionadas por grupo como

$$p_g := \Pr_Q(f(X) = 1 \mid A = g), \qquad g \in \{a, b\}. \tag{1}$$

La disparidad SP es $g_Q^{\mathrm{SP}}(f) := |p_a - p_b|$, y la prueba de cumplimiento basada en tolerancia se escribe como

$$H_0^{\mathrm{SP}} : |p_a - p_b| \leq \delta \qquad \text{frente a} \quad H_1^{\mathrm{SP}} : |p_a - p_b| > \delta. \tag{2}$$

Igualdad de oportunidades (EO) [7] EO condiciona sobre la etiqueta verdadera $Y = 1$. Definimos

$$q_g := \Pr_Q(f(X) = 1 \mid A = g, Y = 1), \qquad g \in \{a, b\}. \tag{3}$$

La disparidad EO es $g_Q^{\mathrm{EO}}(f) := |q_a - q_b|$, y la prueba de cumplimiento se expresa como

$$H_0^{\mathrm{EO}} : |q_a - q_b| \leq \delta \qquad \text{frente a} \qquad H_1^{\mathrm{EO}} : |q_a - q_b| > \delta. \tag{4}$$

2.2. Modelos Secuenciales de Verosimilitud

Bajo SP, el auditor muestrea uniformemente sin reemplazo del conjunto completo; bajo EO, muestrea sobre el subconjunto filtrado $Y = 1$, donde cada consulta es informativa por construcción. La secuencia de consultas de EO se genera mediante una permutación aleatoria uniforme de este conjunto filtrado. Sea $\hat{Y}_t = f(X_t)$ la predicción observada en el paso t. En este escenario de conjunto finito, las verosimilitudes de Bernoulli presentadas a continuación se utilizan como modelos de trabajo operativos para la acumulación secuencial de evidencia sobre las observaciones reveladas. Aunque estas expresiones son exactas para un muestreo de Bernoulli independiente, aquí sirven como modelos manejables de estilo GLR bajo consultas sin reemplazo.

SP (grupos binarios): Para SP, $E_t^{\mathrm{SP}} = \{(A_i, \hat{Y}_i) : i = 1, ..., t\}$ denota los datos disponibles en el instante t. Bajo el modelo de verosimilitud de trabajo, condicionado a los atributos sensibles consultados $\{A_i\}_{i=1}^t$, los resultados predichos $\{\hat{Y}_i\}_{i=1}^t$ se modelan como variables de Bernoulli condicionalmente independientes con probabilidades de éxito específicas por grupo, es decir, $\hat{Y}_i \mid (A_i = g) \sim \mathrm{Bernoulli}(p_g)$ para $g \in \{a, b\}$. Sean $n_g(t)$ y $s_g(t)$ los recuentos muestrales del grupo y los recuentos de resultados positivos hasta el tiempo t; por brevedad, escribimos $n_g = n_g(t)$ y $s_g = s_g(t)$. Entonces, la verosimilitud condicional en el tiempo t es

$$L_t^{\mathrm{SP}}(p_a, p_b) = p_a^{s_a}(1 - p_a)^{n_a - s_a}\, p_b^{s_b}(1 - p_b)^{n_b - s_b}. \tag{5}$$

EO (grupos binarios): Bajo EO, la inferencia es condicional a $Y = 1$ y, por lo tanto, se formula sobre la distribución de referencia condicional Q_+. Los datos se definen como $E_t^{\mathrm{EO}} = \{(A_i, \hat{Y}_i) : i = 1, ..., t\}$ con $(X_i, A_i) \sim Q_+$. Condicionado a los atributos sensibles consultados, los resultados predichos de EO para $g \in \{a, b\}$ satisfacen $\hat{Y}_i \mid (A_i = g) \sim \text{Bernoulli}(q_g)$. Sean $m_g(t)$ y $r_g(t)$ los recuentos análogos sobre este conjunto; de nuevo, por brevedad escribimos $m_g = m_g(t)$ y $r_g = r_g(t)$. Así, la verosimilitud condicional para EO es

$$L_t^{\mathrm{EO}}(q_a, q_b) = q_a^{r_a}(1 - q_a)^{\,m_a - r_a}\, q_b^{r_b}(1 - q_b)^{\,m_b - r_b}. \tag{6}$$

2.3. Estadísticos de Auditoría bajo Diferentes Regímenes de Acceso

La información disponible para un auditor depende del nivel de acceso proporcionado por el modelo desplegado. Consideramos tres regímenes de acceso.

Acceso únicamente a decisiones (solo predicciones): Bajo el acceso únicamente a decisiones (a menudo considerado acceso de caja negra), el auditor observa únicamente la predicción binaria $\hat{Y} = f(X) \in \{0, 1\}$ devuelta por el modelo. Este régimen coincide directamente con las definiciones de SP y EO basadas en decisiones de la Sección 2.1. En este régimen, los resultados para el grupo $g \in \{a, b\}$ se modelan como variables de Bernoulli. Por lo tanto, el estadístico GLR secuencial utiliza la verosimilitud de Bernoulli descrita en la sección anterior, donde los estadísticos suficientes son el número de muestras y el número de resultados positivos observados para cada grupo bajo SP y EO, respectivamente.

Acceso a puntuaciones (scores): En muchos sistemas desplegados, el modelo expone adicionalmente puntuaciones de predicción o probabilidades $S(x) \in [0, 1]$, que corresponden típicamente a la salida sigmoide de un clasificador. Bajo este acceso de caja gris, consideramos auditorías *proxy* basadas en las medias de las puntuaciones condicionales al grupo. Para la auditoría *proxy* de tipo SP, para cada $g \in \{a, b\}$ definimos $\mu_g^{(S)} := \mathbb{E}[S(X) \mid A = g]$ y la disparidad *proxy* correspondiente como $d_{\mathrm{SP}}^{(S)} := |\mu_a^{(S)} - \mu_b^{(S)}|$. Para la de tipo EO, definimos $\mu_{g,+}^{(S)} := \mathbb{E}[S(X) \mid A = g, Y = 1]$, con una disparidad *proxy* $d_{\mathrm{EO}}^{(S)} := |\mu_{a,+}^{(S)} - \mu_{b,+}^{(S)}|$. Operacionalmente, modelamos las puntuaciones observadas como variables gaussianas con medias específicas por grupo y varianza común: $S_i \mid A_i = g \sim \mathcal{N}(\mu_g^{(S)}, \sigma_S^2)$, y de forma análoga al condicionar en $Y = 1$ para EO. Dado que estas cantidades basadas en puntuaciones no son idénticas a las decisiones discretas de SP/EO, utilizamos una tolerancia operacional en el espacio de puntuaciones en lugar de identificarla con la tolerancia de decisión discreta δ.

Acceso a logits: En algunos entornos, el auditor puede observar los logits presigmoides $Z(x)$ del modelo en lugar de las probabilidades. Los logits proporcionan una señal continua no acotada que a menudo contiene más información sobre la confianza interna (grado de confianza del clasificador) del clasificador. Bajo este régimen de acceso ampliado a las salidas, que no requiere acceso a los parámetros del modelo, definimos de manera análoga auditorías *proxy* basadas en las medias de los logits condicionales al grupo. Para SP, sea $\mu_g^{(Z)} := \mathbb{E}[Z(X) \mid A = g]$, y para EO, sea $\mu_{g,+}^{(Z)} := \mathbb{E}[Z(X) \mid A = g, Y = 1]$, obteniendo las

disparidades *proxy* $d_{\mathrm{SP}}^{(Z)} := |\mu_a^{(Z)} - \mu_b^{(Z)}|$ y $d_{\mathrm{EO}}^{(Z)} := |\mu_{a,+}^{(Z)} - \mu_{b,+}^{(Z)}|$. Modelamos los logits observados mediante una verosimilitud Gaussiana con medias específicas por grupo y varianza común: $Z_i \mid A_i = g \sim \mathcal{N}(\mu_g^{(Z)}, \sigma_Z^2)$, condicionando de nuevo en $Y = 1$ para EO. Es importante destacar que no identificamos una tolerancia del espacio de probabilidad con una tolerancia única del espacio de logits, ya que la transformación logística es no lineal; en su lugar, se utiliza un umbral operacional en el espacio de logits. Por tanto, las auditorías basadas en puntuaciones y logits deben interpretarse como auditorías proxy de disparidades continuas, y no como pruebas exactas de las restricciones SP/EO definidas sobre decisiones discretas. En todos los regímenes, la regla de parada secuencial conserva la misma forma; lo que cambia es el modelo de observación y, para observables más ricos, la disparidad auditada. Para las auditorías *proxy* basadas en puntuaciones y logits, las verosimilitudes Gaussianas se utilizan igualmente como modelos de trabajo sobre las observaciones del conjunto finito revelado, en lugar de como leyes de muestreo exactas sin reemplazo.

2.4. Prueba Secuencial de la Razón de Verosimilitudes Generalizada

Para llevar a la práctica las auditorías como pruebas estadísticas de hipótesis, adoptamos un marco GLR. Utilizamos GLR porque tanto H_0 como H_1 definen conjuntos de valores paramétricos admisibles, no distribuciones individuales. Para las auditorías basadas únicamente en decisiones, cuyo objetivo es evaluar la disparidad definida sobre decisiones discretas, sea $\theta = (\theta_a, \theta_b)$ el vector de parámetros de Bernoulli específicos de cada grupo. Definimos el funcional de disparidad $d(\theta) := |\theta_a - \theta_b|$. Dado un nivel de tolerancia $\delta > 0$, definimos los conjuntos paramétricos de cumplimiento y violación como $\Theta_0(\delta) = \{\theta \in [0,1]^2 : d(\theta) \leq \delta\}$ y $\Theta_1(\delta) = \{\theta \in [0,1]^2 : d(\theta) > \delta\}$.

En el instante t, sea $L_t(\theta)$ la verosimilitud correspondiente. Por conveniencia numérica, expresamos el estadístico GLR en términos de la log-verosimilitud $\ell_t(\theta) := \log L_t(\theta)$, que se escribe como:

$$\Lambda_t = \sup_{\theta \in \Theta_1(\delta)} \ell_t(\theta) - \sup_{\theta \in \Theta_0(\delta)} \ell_t(\theta). \tag{7}$$

Para las auditorías *proxy* basadas en puntuaciones y logits, se utiliza la misma construcción GLR reemplazando θ por el par correspondiente de medias condicionales al grupo y tomando L_t como la verosimilitud Gaussiana bajo un modelo de varianza común. En ese caso, las hipótesis nula y alternativa se definen mediante la disparidad *proxy* correspondiente en el espacio de puntuaciones o logits, en lugar de la disparidad de decisión discreta de SP/EO. En las pruebas *proxy* gaussianas, la varianza común se trata como un parámetro auxiliar y se perfila en cada paso utilizando la varianza residual combinada bajo el ajuste restringido o no restringido correspondiente.

Después de t consultas de auditoría, el estadístico Λ_t se actualiza secuencialmente y se compara con los límites de parada superior e inferior (u, ℓ). Debido a que este escenario implica una prueba GLR compuesta, los umbrales clásicos de la prueba secuencial de la razón de verosimilitudes ($SPRT$) no se postulan aquí como garantías exactas para muestras finitas. En su lugar, los límites se tratan

como umbrales de decisión operativos para la auditoría GLR secuencial. Para niveles de error nominales (α, β), utilizamos los valores inspirados en la SPRT $u = \log((1-\beta)/\alpha)$ y $\ell = \log(\beta/(1-\alpha))$. Con $\alpha = 0{,}05$ y $\beta = 0{,}2$, esto resulta en $u \approx 2{,}77$ y $\ell \approx -1{,}56$.

En los experimentos, utilizamos un presupuesto nominal máximo de consultas de $B = 4000$ instancias. Los límites (u, ℓ) son los umbrales operativos fijos anteriores y se mantienen constantes en todos los conjuntos de datos, modelos, métricas y regímenes de acceso. Dado que este artículo se centra en el comportamiento comparativo de la auditoría en lugar de en una calibración exacta para muestras finitas, presentamos las frecuencias de decisión empíricas y los tiempos de parada en lugar de afirmar garantías exactas sobre los errores de tipo I y de tipo II. Bajo la implementación con conjunto finito, el límite efectivo de parada es $T_{\text{máx}}^{(m)} = \text{mín}\{B, N_{\text{pool}}^{(m)}\}$, donde $N_{\text{pool}}^{(m)}$ denota el tamaño del conjunto de auditoría para la métrica m. Para SP, $N_{\text{pool}}^{(\text{SP})}$ es el tamaño del conjunto de auditoría retenido completo, mientras que para EO, $N_{\text{pool}}^{(\text{EO})}$ es el tamaño del conjunto filtrado $Y = 1$ y puede ser menor que B.

La auditoría rechaza H_0 si $\Lambda_t \geq u$, acepta H_0 si $\Lambda_t \leq \ell$, y continúa el muestreo en caso contrario. Si no se cruza ningún límite dentro del presupuesto máximo de auditoría B o antes de que se agote el conjunto de auditoría finito correspondiente, el resultado se declara inconcluyente. La maximización restringida se reduce a una optimización unidimensional sobre la frontera de la restricción de tolerancia y se resuelve mediante la optimización numérica de la función de log-verosimilitud, la cual es estrictamente cóncava.

3. Configuración Experimental

3.1. Conjuntos de Datos Utilizados y Modelos

Evaluamos nuestro marco de auditoría en dos conjuntos de datos tabulares de uso habitual en estudios de equidad en estudios de equidad: **American Community Survey (ACS)** y **Adult**. La tarea ACS Income, procedente del conjunto de referencia Folktables [4], predice si los ingresos de un individuo superan los 50 000 dollares basándose en atributos demográficos y de empleo; nos centramos en los datos de California correspondientes a la encuesta anual de 2018, utilizando el **sexo** como atributo sensible con dos grupos (hombre y mujer). El conjunto de datos Adult, del repositorio de la Universidad de California (UCI), comparte el mismo objetivo de predicción (ingresos superiores a \$50,000) y también utiliza el **sexo** como atributo sensible. Ambos se preprocesan con codificación *one-hot* y estandarización numérica. Cada conjunto de datos se divide en particiones de entrenamiento y prueba, y esta última partición reservada sirve como conjunto de auditoría.

Para estudiar cómo depende el comportamiento de la auditoría de la capacidad del modelo y de la estabilidad del entrenamiento, construimos construimos cuatro modelos basados en perceptrones multicapa que posteriormente sometemos a auditoría. Todos los modelos se tratan como sistemas desplegados fijos durante

la auditoría. El auditor no tiene acceso a los parámetros del modelo ni a los datos de entrenamiento.

Arquitecturas base: Todos los modelos base se implementan como redes neuronales de propagación hacia delante (*feedforward*) con activaciones de unidades lineales rectificadas (*Rectified Linear Units*, ReLU) y una capa de salida con activación sigmoide. El modelo **Robust-Base** consta de dos capas ocultas con 64 y 32 unidades, respectivamente, y se entrena durante 30 épocas usando el optimizador Adam [9] con tasa de aprendizaje 10^{-3} y tasa de *dropout* 0.1. El modelo **Unstable-Base** consta de una única capa oculta con 16 unidades, se entrena durante una época con tasa de aprendizaje 10^{-2} y no utiliza regularización. Esta configuración induce mayor varianza y subajuste, lo que propicia una mayor variabilidad en las regiones de decisión.

Modelos con restricciones de equidad: Para obtener variantes con restricciones de equidad, aplicamos el marco de reducción de gradiente exponenciado con una restricción de paridad demográfica (*Demographic Parity*, DP) [1, 2]. El modelo **Robust-DP** se entrena con parámetro de tolerancia $\epsilon = 0{,}02$ y hasta 10 iteraciones de reducción, mientras que **Unstable-DP** usa $\epsilon = 0{,}05$ y hasta 2 iteraciones de reducción.

3.2. Regímenes de Acceso del Auditor y Criterios de Equidad

Consideramos tres regímenes de observabilidad que determinan qué información del modelo desplegado puede consultar el auditor: (i) **Acceso solo a decisiones**, donde el auditor observa únicamente las predicciones binarias $f(X) \in \{0, 1\}$ y realiza la auditoría de SP/EO basada en decisiones discretas (Sección 2.1); (ii) **Acceso a puntuaciones**, observando las puntuaciones de predicción $p(X)$ para realizar auditorías *proxy* basadas en dichas puntuaciones; y (iii) **Acceso a logits**, observando los logits pre-sigmoides $z(X)$ para ejecutar auditorías *proxy* análogas basadas en logits. Estos regímenes nos permiten estudiar cómo afecta la información adicional del modelo a la eficiencia de la auditoría secuencial bajo un presupuesto nominal máximo de $B = 4000$ consultas y umbrales GLR operativos fijos.

Instanciamos este marco utilizando dos criterios de equidad grupal, evaluando el cumplimiento frente a una tolerancia $\delta > 0$: La **SP**, que exige que $|\Pr(f(X){=}1 \mid A{=}a) - \Pr(f(X){=}1 \mid A{=}b)| \leq \delta$, y la **EO**, que exige que

$$|\Pr(f(X){=}1 \mid Y{=}1, A{=}a) - \Pr(f(X){=}1 \mid Y{=}1, A{=}b)| \leq \delta$$

. Operacionalmente, la prueba secuencial de EO se restringe a un conjunto de auditoría filtrado que contiene únicamente instancias para las que $Y = 1$. A lo largo de los experimentos, las etiquetas de cumplimiento de referencia (oráculo) se calculan siempre a partir de las disparidades de decisión discreta de SP/EO de la Tabla 1. Para el acceso a puntuaciones y logits, los procedimientos secuenciales se interpretan, por tanto, como auditorías *proxy* que preservan la misma estructura de condicionamiento de tipo SP o EO, pero operan sobre puntuaciones o logits en lugar de hacerlo sobre decisiones discretas.

Tabla 1: Disparidades de equidad a nivel poblacional en los conjuntos de prueba.

Dataset	Modelo	Disparidad SP	Disparidad EO
ACS	Robust-DP	0.0287	0.0613
ACS	Unstable-DP	0.0817	0.0011
ACS	Robust-Base	0.0923	0.0202
ACS	Unstable-Base	0.0894	0.0085
Adult	Robust-DP	0.0043	0.3233
Adult	Unstable-DP	0.1424	0.0211
Adult	Robust-Base	0.1716	0.0753
Adult	Unstable-Base	0.1706	0.0819

3.3. Protocolo de Auditoría y Evaluación de Referencia

Todos los experimentos utilizan un muestreo aleatorio uniforme sin reemplazo a partir de un conjunto de auditoría finito y aleatoriamente reordenado, el cual continúa hasta que se cruza un límite de parada o se agota el conjunto disponible. Este diseño nos permite aislar el efecto del régimen de acceso y del estadístico de auditoría sobre el comportamiento de parada, sin introducir efectos de muestreo adaptativo. El conjunto equivale a la partición de prueba completa para SP, y se restringe al subconjunto $Y = 1$ para EO, lo que reduce intrínsecamente el tamaño de la población disponible. En consecuencia, el procedimiento de EO opera sobre una población finita más pequeña y cada consulta de EO es informativa por construcción. Para garantizar la reproducibilidad, el número de consultas en cada auditoría está acotado por $\min(B, N_{\text{pool}})$, donde B es el presupuesto nominal de consultas y N_{pool} es el tamaño del conjunto de auditoría disponible. Para SP, N_{pool} corresponde al conjunto de prueba retenido completo, mientras que para EO corresponde al subconjunto de instancias etiquetadas positivamente. Los recuentos por grupo en estos conjuntos determinan la rapidez con la que el estadístico GLR acumula evidencia.

Para evaluar la corrección de la auditoría *post-hoc*, se calculan las etiquetas de cumplimiento de referencia (oráculo) (cumplimiento frente a violación para δ) usando las disparidades SP y EO exactas de decisiones discretas en toda la población (Tabla 1). Esta información a nivel poblacional permanece oculta para el auditor secuencial. Las métricas de rendimiento se calculan a partir de 20 ejecuciones independientes independientes con diferentes semillas aleatorias, y se reportan la media empírica y la desviación estándar.

4. Resultados numéricos

La Tabla 1 presenta las disparidades de equidad a nivel poblacional para $\delta = 0{,}05$, usadas como referencia para la evaluación *a posteriori*. Con $\delta = 0{,}05$: en ACS, Robust-DP cumple SP pero viola EO; Unstable-DP viola SP pero cumple EO. En Adult, Robust-DP cumple SP pero viola EO de forma clara; Unstable-DP viola SP y cumple EO. Las Tablas 2 y 3 presentan los modelos base; cada configuración se evalúa en 20 ejecuciones independientes.

Tabla 2: Auditoría secuencial bajo SP. La columna SP ref. reporta disparidades SP de decisiones discretas. El acceso solo a decisiones corresponde a la auditoría basada en decisiones discretas, mientras que el acceso a puntuaciones y a *logits* denota auditorías *proxy*. Las entradas reportan la decisión y el número medio de consultas hasta la finalización (± desv. est.). "Rech." (Rechaza).

Dataset - Modelo	**SP ref.**	**Solo decisiones**	**Puntuaciones**	***Logits***
ACS - Rob.-Base	0.0923	Rech., 2349 ± 1345	Rech., 2157 ± 1443	Rech., 708 ± 543
ACS - Unst.-Base	0.0894	Rech., 2412 ± 1587	Rech., 2406 ± 1477	Rech., 691 ± 595
Adult - Rob.-Base	0.1716	Rech., 191 ± 215	Rech., 94 ± 75	Rech., 77 ± 54
Adult - Unst.-Base	0.1706	Rech., 215 ± 178	Rech., 114 ± 83	Rech., 75 ± 50

4.1. Paridad Estadística

Como se muestra en la Tabla 2, las decisiones obtenidas en los distintos regímenes de acceso coinciden, en términos agregados, con las etiquetas de no conformidad calculadas a partir de las decisiones discretas para los modelos base. Para el acceso a puntuaciones y logits, estos resultados deben interpretarse como auditorías *proxy* en relación con la referencia de decisión discreta de SP. Sin embargo, el número de consultas requeridas varía sustancialmente según el nivel de acceso. La auditoría basada únicamente en decisiones requiere sistemáticamente más muestras, el acceso a puntuaciones reduce este coste al explotar información de confianza, y el acceso a logits proporciona las mayores ganancias.

El efecto del acceso es particularmente pronunciado en ACS, donde las disparidades SP de los modelos base están relativamente cerca del umbral de tolerancia. Por ejemplo, en el modelo Robust-Base, el número medio de muestras necesarias disminuye de 2349 bajo el acceso únicamente a decisiones a 708 bajo el acceso a logits. Un patrón similar se observa para el modelo Unstable-Base, donde la media disminuye de 2412 a 691. Por lo tanto, en ACS, el acceso a logits reduce el coste de la auditoría aproximadamente a un tercio en comparación con el acceso únicamente a predicciones.

La misma tendencia cualitativa se mantiene en Adult. En este caso, las violaciones de SP son sustancialmente mayores, por lo que todos los regímenes de acceso finalizan mucho más rápido. Incluso en este escenario más sencillo, disponer de un acceso más rico sigue siendo beneficioso: para Robust-Base, el requerimiento medio de muestras disminuye de 191 bajo el acceso únicamente a decisiones a 94 bajo el acceso a puntuaciones y a 77 bajo el acceso a logits; para Unstable-Base, los valores correspondientes son 215, 114 y 75. Estos resultados indican que las salidas más ricas del modelo mejoran de manera consistente la eficiencia de la auditoría de SP, aunque la ganancia absoluta depende de lo alejada que se encuentre la disparidad de equidad real respecto de la frontera normativa. La Figura 1 visualiza estas tendencias para ambos conjuntos de datos.

4.2. Igualdad de Oportunidades

En comparación con SP, la auditoría de EO es sustancialmente más desafiante, como se muestra en la Tabla 3. Esto se debe a que EO se lleva a cabo directamente sobre el conjunto de auditoría condicionado más pequeño, restringido a $Y = 1$,

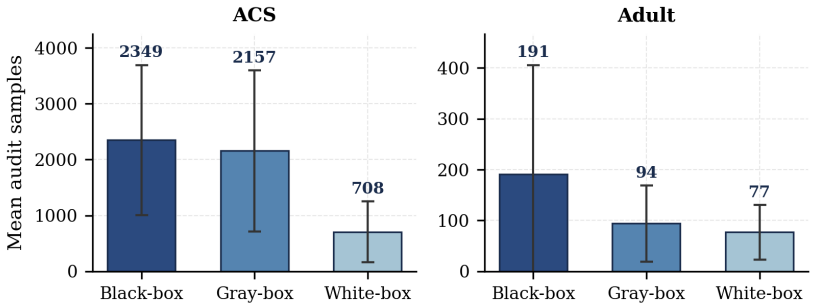


Figura 1: Coste de auditoría secuencial bajo SP en los distintos regímenes de acceso para los conjuntos de datos ACS y Adult. Los paneles se muestran por separado. Las barras de error denotan una desviación estándar calculada a partir de 20 ejecuciones. Un acceso más rico reduce el coste en consultas en ambos conjuntos de datos.

Tabla 3: Auditoría secuencial bajo EO. Los valores EO ref. son disparidades EO de decisiones discretas. El acceso solo a decisiones denota la auditoría basada en decisiones discretas; las columnas de acceso a puntuaciones y a logits denotan auditorías *proxy* condicionadas en $Y = 1$. Las entradas reportan la decisión y el número medio de consultas ($\pm$ desv. est.). "Inc." (inconcluyente) indica que no se cruza ninguna frontera dentro del presupuesto o del conjunto EO.

Dataset - Modelo	**EO ref.**	**Solo decisiones**	**Puntuaciones**	***Logits***
ACS - Rob.-Base	0.0202	Acpt., 2711 ± 1123	Acpt., 1060 ± 1098	Inc., 3329 ± 1214
ACS - Unst.-Base	0.0085	Acpt., 1257 ± 702	Acpt., 550 ± 375	Inc., 2829 ± 1487
Adult - Rob.-Base	0.0753	Inc., 2751 ± 1224	Inc., 2757 ± 1220	Inc., 2512 ± 1334
Adult - Unst.-Base	0.0819	Inc., 2492 ± 1340	Inc., 3045 ± 952	Inc., 2717 ± 1171

lo que limita el número total de consultas disponibles y puede generar recuentos grupales menos favorables para la acumulación de evidencia.

En ACS, ambos modelos base cumplen el criterio de EO, y las auditorías con acceso solo a decisiones y a puntuaciones concluyen que ambos modelos cumplen el criterio en la mayoría de las ejecuciones, pero con un coste de consultas mayor que bajo SP; por su parte, el acceso a logits frecuentemente no alcanza una decisión concluyente. En Adult, ambos modelos violan EO solo moderadamente (disparidades de referencia de 0,0753 y 0,0819), y los tres niveles de acceso terminan con frecuencia con un resultado inconcluyente, con recuentos medios de muestras que oscilan aproximadamente entre 2492 y 3045, y sin una ganancia de eficiencia consistente derivada de un acceso más rico.

En conjunto, estos resultados demuestran que las ganancias asociadas al régimen de acceso dependen de la métrica utilizada. Mientras que un acceso más rico mejora de manera consistente la eficiencia de la auditoría de SP, EO sigue siendo intrínsecamente más difícil debido al menor tamaño del conjunto condicionado. Por lo tanto, los resultados de EO bajo acceso a puntuaciones y logits deben interpretarse con cautela como un resultados de las auditorías*proxy*, en lugar de como evidencia de que un acceso más rico mejora de manera uniforme la auditoría de equidad.

4.3. Auditoría Secuencial Frente Tamaño Muestral Fijo

La Tabla 4 compara el auditor GLR secuencial con una línea base con tamaño muestral fijo (*fixed-sample baseline*) que utiliza el presupuesto completo de consultas. Por motivos de comparabilidad, esta tabla reporta únicamente la auditoría basada en decisiones discretas, correspondiente al régimen de acceso solo a decisiones. El cálculo de la exactitud (*accuracy*) trata los resultados secuenciales inconcluyentes como incorrectos en relación con la etiqueta de referencia. La línea base con tamaño muestral fijo evalúa la misma disparidad de decisión discreta tras agotar todo el presupuesto disponible, utilizando B consultas para SP y $\min(B, N_{Y=1})$ consultas para EO. En ambos conjuntos de datos, la auditoría secuencial reduce sustancialmente el número de consultas requeridas en comparación con el enfoque con tamaño muestral fijo. Esta reducción es particularmente pronunciada en el conjunto de datos Adult bajo SP, donde las decisiones se obtienen con menos de 250 muestras de media, en comparación con el presupuesto completo de 4000.

Sin embargo, esta eficiencia se acompaña de una reducción en la exactitud de la decisión en los escenarios más desafiantes. En particular, bajo EO en el conjunto de datos Adult, el auditor secuencial alcanza una exactitud considerablemente menor que la línea base con tamaño muestral fijo, lo que refleja la dificultad de los casos cercanos al umbral y los tamaños de muestra efectivos limitados.

Estos resultados resaltan una relación de compromiso (*trade-off*) fundamental entre eficiencia y exactitud: la auditoría secuencial puede reducir drásticamente el coste de las consultas, pero puede requerir presupuestos más grandes o ejecuciones repetidas para igualar la fiabilidad de las pruebas con tamaño muestral fijo.

5. Discusión y Trabajo Futuro

Nuestros resultados muestran que la eficiencia de la auditoría depende del régimen de acceso: las salidas más ricas (puntuaciones/logits) reducen el coste en consultas bajo SP, pero actúan como *proxies* en lugar de como pruebas exactas

Tabla 4: Auditoría secuencial frente a auditoría con tamaño muestral fijo para el auditor de decisiones discretas con acceso solo a decisiones. La exactitud (*accuracy*) se mide respecto a la decisión de referencia, contando los resultados inconcluyentes como incorrectos. Para EO, el número de muestras del método de tamaño fijo es $\min(B, N_{Y=1})$, que equivale a 3362 para Adult.

		Secuencial		**Fijo**	
Dataset - Modelo	**Métrica**	**Exactitud**	**Muestras**	**Exactitud**	**Muestras**
ACS – Robust-Base	SP	0.75	2349 ± 1345	1.00	4000
ACS – Robust-Base	EO	0.70	2711 ± 1123	0.95	4000
ACS – Unstable-Base	SP	0.70	2412 ± 1587	0.95	4000
ACS – Unstable-Base	EO	0.90	1257 ± 702	1.00	4000
Adult – Robust-Base	SP	1.00	191 ± 215	1.00	4000
Adult – Robust-Base	EO	0.20	2751 ± 1224	1.00	3362
Adult – Unstable-Base	SP	1.00	215 ± 178	1.00	4000
Adult – Unstable-Base	EO	0.30	2492 ± 1340	1.00	3362

para los criterios basados en decisiones. Además, la evaluación con tolerancia explícita introduce una ambigüedad inevitable cerca de los límites normativos, lo que produce resultados inconcluyentes dentro de presupuestos de consulta prácticos. Finalmente, EO sigue siendo intrínsecamente más difícil debido al menor tamaño del conjunto condicionado. Aunque este estudio establece un marco secuencial para la clasificación binaria, las extensiones a escenarios multigrupo o multiclase quedan abiertas. Las principales limitaciones incluyen la naturaleza *proxy* de las métricas de puntuaciónes/logits y la calibración operativa, en lugar de exacta, de los umbrales GLR. El trabajo futuro explorará estrategias de consulta adaptativas y secuencias de confianza uniformes en el tiempo.

Agradecimientos

Este trabajo ha sido financiado por el Consejo Europeo de Investigación (ERC) en el marco de Horizonte 2020 (Acuerdo de subvención n.º 101003431).

Referencias

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Proceedings of the 35th International Conference on Machine Learning. pp. 60–69 (2018)
2. Bird, S., Dudík, M., Edgar, R., et al.: Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Technical Report (2020)
3. Cen, S.H., Alur, R.: From transparency to accountability and back: A discussion of access and evidence in AI auditing. In: Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (2024)
4. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring Adult: New datasets for fair machine learning. In: Advances in Neural Information Processing Systems Datasets and Benchmarks Track (2021)
5. European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Official Journal of the European Union (2024)
6. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268 (2015)
7. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
8. Howard, S.R., Ramdas, A., McAuliffe, J., Sekhon, J.: Time-uniform, nonparametric, nonasymptotic confidence sequences. The Annals of Statistics **49**(2) (2021)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
10. New York City Council: New York City Local Law 144 of 2021: Automated employment decision tools law. NYC Department of Consumer and Worker Protection (2023)
11. Wald, A.: Sequential Analysis. Wiley (1947)