Núclio Digital School

Máster en Data Analytics

TRABAJO FINAL DE MÁSTER

Análisis Avanzado de Datos para el sistema judicial de EEUU

Análisis de Sesgos en el Algoritmo COMPAS

Autores:

Azahara Bravo Montalbán

Daniel Álvarez

María Fernanda Loza Serrano

Tutor/a:

Cristian Diez

Fecha de entrega:
8 de octubre de 2025

Defensa:
8 de octubre de 2025

ÍNDICE

1. INTRODUCCIÓN

1.1. CONTEXTO Y MOTIVACIÓN DEL PROYECTO

El presente trabajo se desarrolla en el marco del Trabajo Final de Máster en Data Analytics de Nuclio Digital School, dentro de las temáticas propuestas para aplicar de forma práctica las competencias adquiridas durante el programa.

De entre las cuatro opciones disponibles, nuestro grupo eligió el análisis del algoritmo **COMPAS** y la predicción de reincidencia en el sistema judicial de Estados Unidos. Consideramos que se trata de la propuesta con mayor relevancia social y ética, ya que aborda un tema de gran impacto en la vida de las personas: el uso de algoritmos de riesgo para tomar decisiones judiciales.

Nuestra elección está también relacionada con el bagaje personal y profesional del equipo. Dos de los integrantes procedemos del ámbito sanitario y social, donde hemos trabajado en la atención directa a personas y en la gestión de recursos para colectivos vulnerables. Este trasfondo nos ha permitido reconocer de inmediato la importancia de evaluar críticamente sistemas que pueden condicionar el acceso a derechos fundamentales, como la libertad. Por otro lado, la tercera integrante comparte un fuerte compromiso con la justicia social y la equidad, lo que ha aportado una mirada complementaria y enriquecedora al análisis.

En conjunto, nuestro equipo percibió este proyecto no solo como una oportunidad académica para aplicar técnicas de análisis de datos y modelado predictivo, sino también como una forma de contribuir a un debate social y ético de gran actualidad: el papel de la inteligencia artificial y los algoritmos en contextos sensibles como el sistema judicial.

1.2. OBJETIVOS PRINCIPALES

El objetivo general de este trabajo es **analizar críticamente el algoritmo COMPAS**, utilizado en el sistema judicial de EE.UU. para predecir el riesgo de reincidencia, y evaluar sus posibles sesgos y limitaciones desde una perspectiva de Data Analytics.

De manera más específica, nos planteamos los siguientes objetivos:

1. **Aplicar las competencias adquiridas durante el máster** en un caso real de análisis de datos, desarrollando un proyecto que integre limpieza, exploración, modelado y visualización de datos.

2. **Realizar un análisis exploratorio de los datos (EDA)**, identificando patrones, distribuciones y posibles problemas de calidad en las variables demográficas, judiciales y de reincidencia.

3. **Detectar y cuantificar sesgos** en el algoritmo COMPAS, con especial atención a diferencias por etnia y género, y discutir su relevancia ética y social.

4. **Evaluar la capacidad predictiva de COMPAS**, midiendo su precisión y tasas de error (falsos positivos y falsos negativos), y comparando los resultados entre diferentes grupos poblacionales.

5. **Desarrollar un modelo predictivo alternativo**, utilizando un modelo de regresión logística, que sirva como punto de referencia para contrastar la fiabilidad de COMPAS y explorar opciones menos sesgadas.

6. **Diseñar un dashboard interactivo** que integre los resultados más relevantes del análisis, facilitando la comprensión y comunicación de los hallazgos tanto a un público técnico como no técnico.

7. **Extraer conclusiones y formular recomendaciones al sistema judicial estadounidense**, aportando una visión crítica sobre el uso de algoritmos predictivos en la toma de decisiones judiciales y sugiriendo estrategias para mitigar sesgos en futuros desarrollos.

1.3 ALCANCE DEL TRABAJO

El presente proyecto se centra en el análisis de los datos de reincidencia del condado de Broward (Florida, EE.UU.) publicados por ProPublica y en la evaluación crítica del algoritmo COMPAS como herramienta predictiva.

Nuestro trabajo abarca:

- La exploración y limpieza de los datos originales.
- El análisis descriptivo de variables demográficas y judiciales.
- La identificación de sesgos en las puntuaciones de COMPAS, especialmente en relación con etnia y género.
- La evaluación de la capacidad predictiva de COMPAS frente a la reincidencia real.
- La construcción de un modelo alternativo de referencia (regresión logística).
- La comunicación de los resultados mediante visualizaciones en ppt y un dashboard interactivo.
- La formulación de recomendaciones al sistema judicial estadounidense.

Quedan fuera del alcance de este proyecto:

- El desarrollo de modelos predictivos avanzados más allá de la regresión logística (ej. redes neuronales, random forests).
- La validación de los resultados en otros contextos judiciales distintos al condado de Broward.
- El diseño de políticas públicas, que excede nuestras competencias técnicas, aunque sí incluimos recomendaciones generales basadas en los hallazgos.

---

2. DESCRIPCIÓN DE LOS DATOS

2.1. FUENTE DE LOS DATOS

Los datos utilizados en este proyecto provienen de la investigación realizada por **ProPublica** en 2016, en la que se evaluó el uso del algoritmo **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) en el condado de **Broward, Florida (EE.UU.)**. ProPublica obtuvo un conjunto de datos de acusados en espera de juicio (pretrial defendants) y personas en libertad condicional (probationers) del condado mencionado, que habían sido evaluados con el sistema de cribado COMPAS entre el 1 de enero de 2013 y el 31 de diciembre de 2014.

La base de datos que obtuvieron contiene información de más de **11.000 acusados** evaluados en ese lapso de 2 años, para los cuales se registraron tanto las puntuaciones de riesgo asignadas por **COMPAS** como el historial de reincidencia observado durante los dos años posteriores a su primera detención.

En nuestro proyecto se nos ofreció analizar dos datasets “Compas_scores_raw” y “compas_scores_two_years”. Estos archivos son los que más habitualmente han sido utilizados por otros investigadores. Son archivos trabajados por ProPublica basados en los que aportó el condado de Broward y a los que crearon varias columnas para el análisis.

- **Archivo “compas_scores_raw.csv”:** contiene el conjunto completo de acusados en espera de juicio obtenido de la Oficina del Sheriff del Condado de Broward. Este archivo incluye 11757 personas y fué reducido a 10331 personas por ProPublica.
- **Archivo “compas_scores_two_years.csv”**: Este es creado por ProPublica con el propósito de estudiar la reincidencia general a dos años que es diferente al de reincidencia violenta que es más acotada.

Cabe decir que las puntuaciones de riesgo de reincidencia de COMPAS se basan en las respuestas del acusado al cuestionario de cribado COMPAS. ***Esta encuesta es completada por los servicios previos al juicio en cooperación con el acusado, después de su arresto***. En estos datos de ProPublica, esta encuesta, **se administra, por lo general, el mismo día o el día después de que la persona ingresa en la cárcel.**

La idea de predicción de COMPAS abarca distintas dimensiones del riesgo:

- Riesgo de reincidencia.
- Riesgo de reincidencia violenta.
- Riesgo de no comparecencia.

2.2. VARIABLES DISPONIBLES Y SIGNIFICADO

La base de datos incluye distintos tipos de variables:

- **Datos sociodemográficos**: edad, rango de edad, género, etnia.
- **Datos judiciales**: cargos iniciales, fechas de arresto y encarcelamiento, duración de custodia, grado del delito y descripciones de delitos.

- **Puntuaciones COMPAS**: valores de 1 a 10 que reflejan el riesgo calculado por el algoritmo. COMPAS clasifica sus resultados en tres grupos:
    - 1–4: Riesgo bajo.
    - 5–7: Riesgo medio.
    - 8–10: Riesgo alto.

- **Resultado de reincidencia general:** la fecha de recopilación de datos indica si se ha reincidido o no sin límite de tiempo.
- **Resultado de reincidencia**: indicador binario de si la persona fue arrestada de nuevo en los dos años posteriores a la primera detención, creada por ProPublica para acotar el margen de inclusión en su estudio a dos años.

ProPublica recopiló los datos para su investigación y creó una base de datos en Python. A partir de la cual construyó varios subconjuntos de datos en los que fusionó y calculó diversas variables clave como "resultado de reincidencia" (a dos años) y el "período de tiempo entre arrestos".

En el proceso de trabajo, realizamos un **renombrado y adaptación de las variables al castellano** para facilitar la interpretación en los análisis y en las visualizaciones (ejemplo: `decile_score` → `resultado_compas`, `two_year_recid` → `reincide`).

Tabla 1. Descripción de las columnas el dataset original

Nombre Columna	Descripción	Columna clave
id	Identificador único del individuo	✔️
name	Nombre completo del individuo	
first	Primer nombre del individuo	
last	Apellido del individuo	
compas_screening_date	Fecha de la evaluación COMPAS	
sex	Género del individuo (Male/Female)	✔️
dob	Fecha de nacimiento del individuo	
age	Edad en años del individuo	✔️
age_cat	Categoría de edad (agrupación)	✔️
race	Grupo racial del individuo	✔️
decile_score	Puntuación COMPAS de reincidencia general (1–10)	✔️
juv_fel_count	Número de delitos juveniles graves (felonies)	✔️
juv_misd_count	Número de delitos juveniles menores (misdemeanors)	✔️
juv_other_count	Otros delitos juveniles	✔️
priors_count	Número total de delitos previos adulto	✔️
days_b_screening_arrest	Días entre el arresto y la evaluación COMPAS	
c_jail_in	Fecha de entrada en la cárcel por el delito actual	
c_jail_out	Fecha de salida de la cárcel por el delito actual	

c_case_number	Número de caso del delito actual	
c_offense_date	Fecha en que ocurrió el delito actual	
c_arrest_date	Fecha de arresto por el delito actual	
c_days_from_compas	Días transcurridos entre el delito actual y la evaluación COMPAS	
c_charge_degree	Grado del delito actual (F: felony, M: misdemeanor)	✔️
c_charge_desc	Descripción del delito actual	
is_recid	Indica si el individuo reincidió en algún momento (booleano)	✔️
r_case_number	Número de caso en la reincidencia	
r_charge_degree	Grado del delito en caso de reincidencia	✔️
r_days_from_arrest	Días entre el arresto actual y el siguiente arresto	
r_offense_date	Fecha del delito reincidente	
r_charge_desc	Descripción del delito en caso de reincidencia	
r_jail_in	Fecha de entrada en la cárcel por reincidencia	
r_jail_out	Fecha de salida de la cárcel por reincidencia	
violent_recid	Indicador de reincidencia en delitos violentos (puede contener nulos)	
is_violent_recid	Indica si reincidió en un delito violento (booleano)	
vr_case_number	Número de caso del delito reincidente violento	
vr_charge_degree	Grado del delito violento reincidente	
vr_offense_date	Fecha del delito violento reincidente	
vr_charge_desc	Descripción del delito violento reincidente	
type_of_assessment	Tipo de evaluación realizada por COMPAS	
decile_score.1	Duplicado de decile_score (puede eliminarse)	
score_text	Categoría de riesgo asociada al score (Low, Medium, High)	✔️
screening_date	Fecha de la evaluación (equivalente a compas_screening_date)	✔️
v_type_of_assessment	Tipo de evaluación de violencia	
v_decile_score	Puntuación COMPAS de violencia (1–10)	
v_score_text	Categoría de riesgo de violencia (Low, Medium, High)	
v_screening_date	Fecha de la evaluación de violencia	
in_custody	Fecha de entrada en custodia preventiva	

out_custody	Fecha de salida de custodia preventiva	
priors_count.1	Duplicado de priors_count (puede eliminarse)	
start	Fecha de inicio del periodo de seguimiento	
end	Fecha de fin del periodo de seguimiento	
event	Estado del evento de seguimiento (si reincidió durante el periodo)	
two_years_recid	Reincidencia en los dos años siguientes al delito original	✔️

2.3. LIMITACIONES Y CALIDAD DE LOS DATOS

El conjunto de datos presenta varias limitaciones que han sido consideradas en el análisis:

- **Valores nulos**: Existen columnas con alta proporción de valores faltantes (por ejemplo, variables relacionadas con reincidencia violenta, con más del 80% de nulos). Evidentemente esa columna no era valorable en nuestro análisis.
- **Cobertura temporal restringida**: Sólo se dispone de información de acusados en 2013–2014 y reincidencia hasta 2016, lo cual limita la generalización a otros periodos.

- **Ámbito geográfico limitado**: Los datos corresponden únicamente al condado de Broward, y por tanto no reflejan necesariamente la realidad de otros estados o del sistema judicial estadounidense en su conjunto.

- **Calidad y sesgo en la recogida de datos**: Al basarse en registros judiciales y en cuestionarios de COMPAS, los datos pueden estar influidos por sesgos preexistentes en el sistema judicial (por ejemplo, mayor probabilidad de arresto en ciertos grupos étnicos que estén concentrados en un código postal específico).

- **Definición de reincidencia**: La reincidencia se mide únicamente como un nuevo arresto en un plazo de dos años, lo que no necesariamente refleja la comisión de un nuevo delito o su gravedad.

Es importante subrayar que el dataset mide la reincidencia únicamente como “nuevo arresto en un plazo de dos años”, lo que no necesariamente refleja la comisión real de un nuevo delito. Esto introduce un sesgo estructural: la proporcionalidad de reincidencia observada en los datos puede estar más relacionada con los patrones de actuación policial y judicial que con la verdadera proporción de delitos cometidos en cada grupo étnico o de género. Por tanto, al interpretar los resultados debemos ser conscientes de que la base de datos refleja la **proporcionalidad de arrestos**, y no necesariamente la **proporcionalidad real de conductas delictivas**.

Asimismo, investigaciones posteriores (Barenstein, 2019) han señalado que ProPublica aplicó criterios temporales distintos al medir la reincidencia: **se permitió un periodo mayor para observar reincidencia en quienes sí reincidían (hasta el 31/12/2014), mientras que para el resto se aplicó un corte más estricto (1/4/2014)**. Este diseño puede haber **inflado artificialmente la tasa de reincidentes en el dataset que utilizamos**, lo que supone una limitación adicional en la calidad de los datos de partida.

---

3. METODOLOGÍA DE TRABAJO

3.1 FASES DEL ANÁLISIS (EDA, SESGOS, MODELADO, VISUALIZACIÓN)

Nuestro trabajo se estructuró en tres notebooks principales que reflejan las fases metodológicas del proyecto:

1. **Notebook 1 – Análisis Exploratorio de Datos y Sesgos (EDA)**
    - Carga y exploración inicial de los datasets de ProPublica (raw y two-years).
    - Limpieza y transformación de variables (renombrado a castellano, control de nulos, duplicados y outliers).
    - Estadísticas descriptivas y distribuciones por etnia, género y rango de edad.
    - Identificación preliminar de sesgos: diferencias en puntuaciones COMPAS, tasas de reincidencia observadas y correlaciones (Spearman).

2. **Notebook 2 – Evaluación y Regresión Logística**
    - Definición de variable objetivo: `reincide` (nuevo arresto en 2 años).
    - Selección de variables predictoras sociodemográficas y judiciales.
    - Codificación de variables categóricas y preparación del dataset.
    - Implementación de un modelo de regresión logística como alternativa ilustrativa a COMPAS, que permite evidenciar sus limitaciones y los sesgos presentes en los datos de origen.
    - Evaluación comparativa entre COMPAS y el modelo: precisión, recall, TPR/FPR y curva ROC-AUC.
    - Análisis de resultados diferenciados por grupos demográficos.

3. **Notebook 3 – Análisis de la controversia metodológica**
    - Estudio crítico de la construcción del dataset por ProPublica y de las críticas de Barenstein.
    - Generación de un tercer dataset de referencia, incorporando criterios más homogéneos de seguimiento temporal.
    - Síntesis de implicaciones: necesidad de datos de mayor calidad y representatividad para mejorar cualquier sistema de predicción de reincidencia.

3.2. HERRAMIENTAS UTILIZADAS

- **Lenguaje y entorno**: Python en Jupyter Notebooks.
- **Librerías de análisis**: `pandas`, `numpy`, `scipy.stats`.
- **Librerías de visualización**: `matplotlib`, `seaborn`, `plotly.express`, `missingno`.
- **Modelado**: `sklearn` (train_test_split, LogisticRegression, métricas ROC/curva).
- **Gestión de fechas y regex**: `datetime`, `re`.
- **Dashboard**: Google Looker Studio, con KPIs para visualizar diferencias en precisión, TPR/FPR y distribución de riesgos entre grupos.
- **Control de calidad**: comprobaciones propias con funciones auxiliares.

3.3. FUNCIONES AUXILIARES (lib_propias.py)

Se usó una librería propia (`lib_propias.py`) para estandarizar procesos y evitar duplicación de código. Entre las funciones más relevantes:

- **Exploración inicial**:
  - `check_df()`: descripción básica y extendida de la estructura del dataset.
  - `id_valores_problem()`: identificación de nulos, duplicados y outliers.

- **Limpieza**:
  - `procesar_fecha()`: estandarización de formatos de fecha.

- **EDA y visualización**:
  - `graficar_histograma_px()`, `graficar_barras_px()`, `graficar_pie_chart()`: distribuciones univariantes.
  - `graficar_boxplot_bivariable_px()`, `graficar_histograma_bivariable_px()`: comparaciones entre variables categóricas y numéricas.
  - `graficar_correlacion()`, `realizar_correlaciones()`: correlaciones y heatmaps.
  - `realizar_crosstab()`: tablas de contingencia con conteos y porcentajes.

- **Estadística avanzada y tests**:
  - `calcular_t_student()`, `evaluar_p_valor()`: pruebas de hipótesis.

Estas funciones facilitaron la **reproducibilidad**, permitieron mantener un **formato uniforme en las visualizaciones** y agilizaron la obtención de insights en cada fase del análisis.

---

4. ANÁLISIS EXPLORATORIO EDA

4.1. LIMPIEZA Y PROCESAMIENTO

Selección del dataset: se trabajó inicialmente con los dos ficheros originales de ProPublica:

- `compas-scores-raw.csv`: contiene todos los registros iniciales.
- `compas-scores-two-years.csv`: versión ya filtrada por ProPublica, limitada a acusados evaluados entre 2013 y 2014 y con seguimiento de reincidencia a 2 años.

El análisis principal se llevó a cabo con el dataset **TwoYears**, ya que es el que permite evaluar la reincidencia de forma consistente. No obstante, el fichero **Raw** se utilizó para comprobar los errores en los cortes de datos que Barenstein identifica en su estudio, lo que nos permitió validar las limitaciones de la depuración realizada por ProPublica.

La limpieza y el procesamiento se basó en los siguientes puntos:

Fechas → `datetime`

- En *TwoYears*: conversión masiva de columnas de fecha con `pd.to_datetime(..., errors='coerce')` (p. ej. `compas_screening_date`, `dob`, `c_jail_in/out`, `r_jail_in/out`, `c_offense_date`, etc.).
- En *Raw*: `DateOfBirth` y `Screening_Date` a `datetime`.

Campos binarios → `bool`

- En *TwoYears*: `event`, `two_year_recid`, `is_violent_recid`, `is_recid` se convierten a booleano con `.astype(bool)`.

Renombrado de columnas (inglés → castellano)

- Se aplica `rename(columns=nuevos_nombres_twoyears)` / `rename(columns=nuevos_nombres)` para usar nombres en castellano (p. ej. `two_year_recid → reincide`, `decile_score → resultado_compas`, `age_cat → rango_edad`, etc.).

Normalización puntual en el dataset *Raw*

- Sustitución de `African-Am` por `African-American` en `Ethnic_Code_Text` para unificar etiquetas del grupo étnico.

4.2. ESTADÍSTICAS DESCRIPTIVAS

El dataset central sobre el que hemos trabajado es el publicado por ProPublica bajo el nombre compas-scores-two-years.csv. Tras la limpieza inicial, quedó compuesto por 7.214 individuos y 53 variables. Cada fila corresponde a un acusado único, por lo que no existen duplicados en la base.

En lo referente a la completitud de los datos, variables fundamentales para el análisis como el género (sex), la etnia (race), la edad (age y age_cat), la puntuación de riesgo (decile_score y score_text), el historial delictivo adulto (priors_count) y el indicador de reincidencia a dos años (two_year_recid) no presentan valores nulos. Esto nos asegura que las comparaciones que realizamos en apartados posteriores no se ven distorsionadas por registros incompletos en estas dimensiones clave.

En cambio, encontramos 307 casos sin información en days_b_screening_arrest (4,26 %) y un alto grado de ausencia en las variables de reincidencia violenta (por ejemplo, violent_recid está vacío en el 100 % de los registros y las variables vr_* superan el 88 % de nulos). Por este motivo, decidimos prescindir de estas últimas en el análisis, al no aportar fiabilidad suficiente.

En cuanto al perfil demográfico, observamos que la muestra está dominada por individuos de 25 a 45 años, que representan el 57 % del total (4.109 personas). Los grupos de jóvenes menores de 25 años y mayores de 45 tienen un peso similar, con 21,2 % y 21,8 % respectivamente. En términos de género, se confirma un claro desequilibrio: el 80,7 % son hombres (5.819) frente a solo 19,3 % de mujeres (1.395), lo que anticipa que cualquier modelo entrenado sobre estos datos podría tener un sesgo implícito hacia patrones masculinos. En la dimensión étnica, la mayoría de la población evaluada es African-American (51,2 %; 3.696 individuos) o Caucasian (34 %; 2.454 individuos).

Los grupos restantes tienen una presencia mucho más reducida: Hispanic (8,8 %), Other (5,2 %), y porcentajes casi testimoniales de Asian (0,4 %; 32 casos) y Native American (0,2 %; 18 casos). Estas cifras confirman la fuerte asimetría en la representación de las distintas comunidades.
Las variables judiciales muestran también información relevante. El grado del cargo actual (c_charge_degree) está marcado en su mayoría por delitos graves (Felonies, 64,7 %) frente a delitos menores (Misdemeanors, 35,3 %). Esta proporción refuerza la idea de que el dataset se centra principalmente en personas acusadas de crímenes de mayor severidad, lo cual puede influir en la interpretación de los riesgos de reincidencia.

En lo que respecta a la puntuación COMPAS, que varía de 1 a 10, la distribución refleja que más de la mitad de los acusados fueron clasificados como riesgo “Bajo” (54 %; 3.897 casos), mientras que un 26,5 % (1.914 casos) se situaron en la categoría de “Medio” y un 19,4 % (1.403 casos) en la de “Alto”. El análisis detallado por deciles muestra una concentración en los valores bajos: el 20 % de la muestra tiene score 1 y los valores van decreciendo hasta llegar al 5,3 % en score 10. Este patrón sugiere que, aunque COMPAS asigna puntuaciones en todo el rango, en la práctica gran parte de la población evaluada es etiquetada en los niveles inferiores.

En cuanto a los antecedentes penales adultos (priors_count), la media es de 3,47 y la mediana de 2, lo que implica que la mayoría de los individuos tienen pocos antecedentes, pero existen casos de reincidencia crónica que elevan el promedio hasta un máximo de 38 antecedentes. Al agrupar por tramos, vemos que casi un tercio de los acusados (29,8 %) no tiene antecedentes, mientras que aproximadamente la mitad (49,1 %) cuenta con entre 1 y 5 registros previos. El 21 % restante agrupa a quienes tienen más de 6 antecedentes, incluidos 626 casos (8,7 %) con entre 11 y 38. Esta distribución confirma que, aunque existe un grupo significativo de personas sin historial, el dataset contiene también perfiles altamente reincidentes que amplían la variabilidad.

Finalmente, la variable “días entre arresto y evaluación COMPAS” (days_b_screening_arrest) merece una atención especial. Entre los 6.907 registros no nulos, la media es de 3,30 días, aunque la mediana de −1 revela que en la mayoría de los casos la evaluación se hizo el mismo día del arresto o justo después. Sin embargo, la dispersión es muy elevada (DE = 75,81) y el rango incluye valores poco plausibles, desde −414 hasta 1.057 días. Detectamos 735 casos (10,2 %) fuera del margen razonable de ±30 días, además de 307 nulos (4,26 %).

Debido a esta anomalía, consideramos relevante realizar un análisis adicional en el Notebook 3, en el que se eliminaron los registros con valores extremos para evaluar cómo afectan a la solidez y validez de los resultados obtenidos.

4.3. DISTRIBUCIÓN DE VARIABLES DEMOGRÁFICAS Y PUNTAJE COMPAS

En este apartado analizamos cómo varían las puntuaciones de COMPAS en función de la **etnia**, el **género** y la **edad**, así como la relación con los **antecedentes penales** de las personas evaluadas.

Pero inicialmente vemos importante entender la representación de las diferentes variables demográficas en los datos.

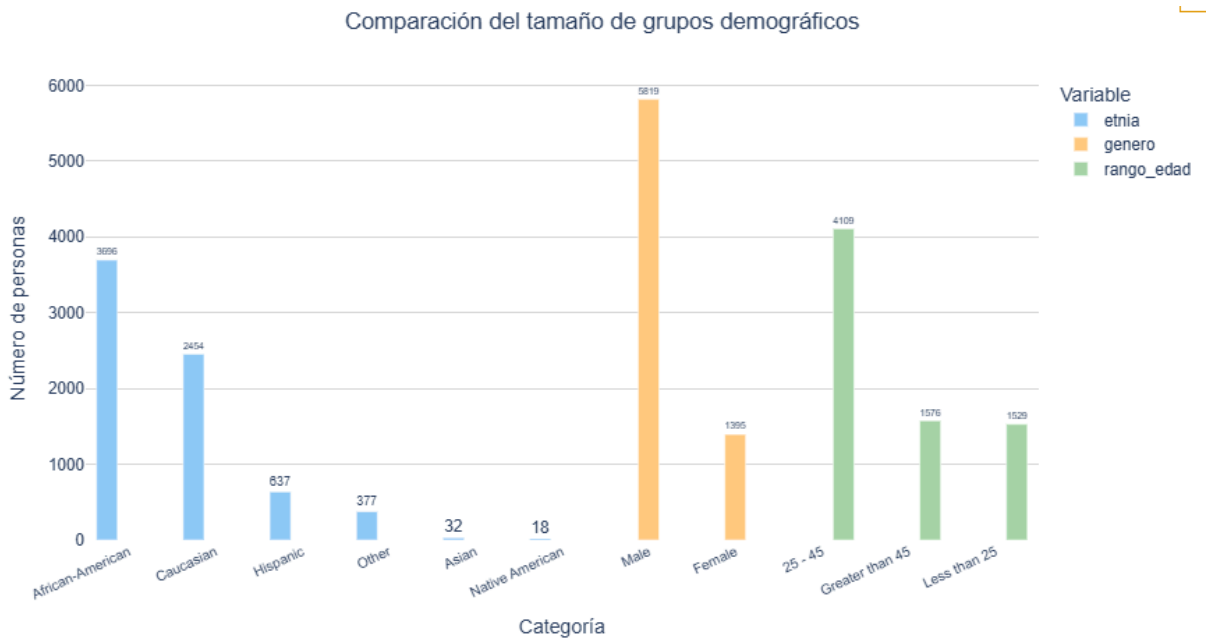


Fig.1 Gráfico de barras de las variables demográficas y su representación en el dataset.

El dataset presenta **fuertes desequilibrios demográficos** que condicionan cualquier análisis.

- **Género**: hay una clara subrepresentación femenina (19%), lo que implica que los modelos se ajustarán mejor a patrones masculinos y serán menos fiables para mujeres.
- **Etnia**: la población está dominada por African-American (51%) y Caucasian (34%), mientras que otros grupos como Asian (0,4%) y Native American (0,2%) apenas tienen casos, lo que limita la validez estadística en ellos.
- **Edad**: la mayoría de los individuos se concentran en la franja 25–45 años, con mucha menor presencia de jóvenes y mayores.

Es importante señalar que los desequilibrios observados no solo reflejan características de la muestra, sino también posibles sesgos estructurales en los procesos de detención en EE.UU., lo que condiciona de partida los datos sobre los que COMPAS realiza sus evaluaciones.

4.3.1 Score COMPAS por grupo demográfico

La tabla siguiente resume la media y la desviación estándar de la puntuación COMPAS por grupo:

Tabla 2: Media de la puntuación ofrecida por Compas según variables demográficas.

Grupo	Categoría	Media score COMPAS	Desviación estándar / Comentario
Etnia	African-American	5.37	Alta variabilidad
	Caucasian	3.74	Media

	Hispanic	3.46	Media
	Other	2.95	Media-baja
	Asian	2.94	Baja
	Native American	6.17*	Muy alta (muestra pequeña)
Género	Hombres	4.59	Más dispersión de valores
	Mujeres	4.17	Menor variabilidad
Edad	<25 años	5.91	Grupo con mayor puntuación
	25-45 años	Valores intermedios	—
	>45 años	2.92	Media más baja

*El valor de Native American debe interpretarse con cautela, ya que la muestra es muy reducida.

Interpretación: La media de puntuación COMPAS varía de manera significativa entre los distintos grupos demográficos. En el plano étnico, los acusados African-American presentan la media más elevada (5,37), lo que sugiere que reciben evaluaciones de mayor riesgo en comparación con los Caucasian (3,74) y con otros grupos minoritarios que se sitúan en valores aún más bajos (Asian 2,94; Other 2,95). El caso de los Native American (6,17) debe interpretarse con mucha cautela debido al reducido tamaño de la muestra (18 individuos). En cuanto al género, los hombres obtienen una media superior (4,59) a la de las mujeres (4,17) y además muestran una dispersión más amplia, lo que indica que las puntuaciones masculinas tienden a variar más entre individuos. Por edad, el patrón es igualmente claro: los menores de 25 años alcanzan la media más alta (5,91), mientras que los mayores de 45 años presentan la más baja (2,92), quedando los individuos de 25 a 45 años en una posición intermedia.

En conjunto, esta gráfica ilustra que **COMPAS no distribuye las puntuaciones de manera uniforme**: tiende a asignar valores más altos a jóvenes, hombres y a determinados grupos étnicos, en particular a los African-American. Estas diferencias constituyen una señal de posible sesgo en el algoritmo y condicionan la interpretación de sus resultados posteriores.

4.3.2 Antecedentes penales por grupo étnico

Para contextualizar estos resultados, se analizaron los antecedentes judiciales registrados.

Tabla 3: Patrón y media de antecedentes por etnia.

Grupo	Media de antecedentes totales	Patrón en delitos juveniles
African-American	4.44	Más antecedentes juveniles que otros grupos
Caucasian	2.59	Muy pocos registros juveniles
Hispanic	2.25	Muy pocos registros juveniles
Native American	6.00*	También más delitos en etapa juvenil
Other/Asian	Valores bajos	Casi sin delitos juveniles

*Muestra muy reducida, interpretarse con cautela.

Interpretación: los datos evidencian que algunos grupos (African-American y Native American) presentan más antecedentes tanto en la adultez como en la juventud. Esto puede contribuir a explicar parte de las diferencias en el score COMPAS. Sin embargo, el hecho de que existan correlaciones con los antecedentes no implica necesariamente que el algoritmo sea justo, cuestión que se aborda en el análisis de la reincidencia real (apartado siguiente).

4.3. REINCIDENCIA A 2 AÑOS

En esta sección analizamos la reincidencia real en un plazo de dos años y cómo se distribuye según las principales variables demográficas y el nivel de riesgo asignado por COMPAS.

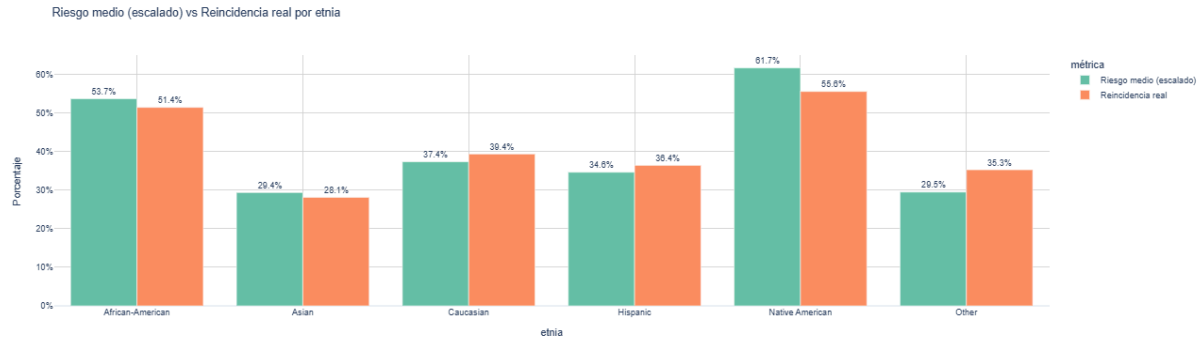
Tabla 4: Porcentajes de reincidencia real por variables demográficas.

Grupo / Categoría	% No reincide	% Reincide
Etnia		
African-American	48.6%	51.4%
Asian	71.9%	28.1%
Caucasian	60.6%	39.4%
Hispanic	63.6%	36.4%
Native American*	44.4%	55.6%
Other	64.7%	35.3%
Género		
Mujer	64.3%	35.7%
Hombre	52.7%	47.3%
Nivel COMPAS		
Bajo	68.8%	31.2%
Medio	46.0%	54.0%
Alto	28.7%	71.3%

* Valores interpretados con cautela por el tamaño reducido de la muestra.

Interpretación:
- Existen diferencias reales de reincidencia entre etnias: African-American y Native American superan el 50%, mientras que Asian presenta la tasa más baja.
- Los hombres reinciden con mayor frecuencia que las mujeres.
- El score COMPAS refleja una gradación clara: a mayor nivel de riesgo, mayor probabilidad de reincidencia (del 31.2% en bajo al 71.3% en alto).

Las siguientes visualizaciones complementan las tablas presentadas y permiten observar de manera más clara las diferencias en las tasas de reincidencia entre grupos. Su valor no está en aportar cifras nuevas, sino en hacer visibles los contrastes de un modo más intuitivo.



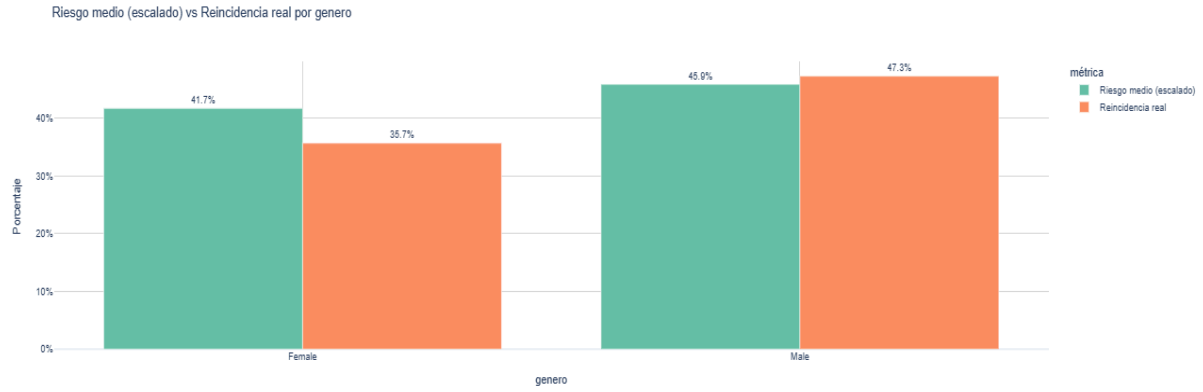


Fig. 2: Gráfico del resultado del riesgo medio escalado resultado de Compas comparado con la reincidencia real por etnia.
Fig. 3: Gráficos Resultado de riesgo medio escalado resultado de Compas comparado con la reincidencia real por género.

Interpretación: La gráfica por etnias muestra de un vistazo que African-American y Native American superan el 50% de reincidencia, mientras que el resto de grupos se sitúan claramente por debajo. En la dimensión de género, la diferencia entre hombres y mujeres se aprecia también con nitidez. Y en la tabla 4, podemos comprobar como la gradación de riesgo COMPAS reproduce la relación esperada: cuanto mayor es el nivel asignado, mayor es la reincidencia real.

Esto confirma cierta capacidad predictiva de COMPAS, pero plantea la necesidad de comprobar si el desempeño es igualmente justo entre los diferentes grupos demográficos.

---

5. IDENTIFICACIÓN DE SESGOS

En esta etapa nos enfocamos en identificar posibles sesgos en la asignación de puntuaciones, con especial atención a los grupos de etnia y género. El objetivo fue determinar si COMPAS tiende a asignar puntuaciones de manera indiscriminada en función de características demográficas, evidenciando patrones de sesgo.

Como punto de partida, elaboramos una síntesis exploratoria focalizada en sesgos, centrada en distribuciones y medidas de tendencia central por grupo para detectar diferencias relevantes en los scores.

A continuación, aplicamos la prueba de Kruskal–Wallis para contrastar si las distribuciones de los scores difirieron significativamente entre múltiples grupos demográficos.

Asimismo, construimos cohortes para igualar condiciones y permitir comparaciones más precisas entre el riesgo asignado por el modelo y la reincidencia observada. Las cohortes se definieron con las variables grado_cargo_compas, rango_antecedentes, rango_edad, etnia y género. Dado que los efectos en etnia y género resultaron más pronunciados, el análisis se centró en estas variables, mientras que rango_edad se incluyó de forma descriptiva como referencia.

A partir de estas cohortes identificamos casos específicos que ilustraron incongruencias entre el riesgo asignado y la reincidencia real, aportando evidencias concretas de posibles sesgos del sistema.

=== COMPAS por ETNIA ===

	etnia	n	media	mediana	sd	p25	p75	minimo	maximo	pct_low	pct_medium	pct_high	reincidencia_real
4	Native American	18	6.17	7.0	2.98	3.25	8.75	2	10	33.3	33.3	33.3	55.6
0	African-American	3696	5.37	5.0	2.83	3.00	8.00	1	10	41.2	31.1	27.7	51.4
2	Caucasian	2454	3.74	3.0	2.60	1.00	5.00	1	10	65.2	23.6	11.2	39.4
3	Hispanic	637	3.46	3.0	2.60	1.00	5.00	1	10	70.2	19.3	10.5	36.4
5	Other	377	2.95	2.0	2.35	1.00	4.00	1	10	79.0	14.1	6.9	35.3
1	Asian	32	2.94	2.0	2.60	1.00	3.50	1	10	75.0	15.6	9.4	28.1

=== COMPAS por GENERO ===

	genero	n	media	mediana	sd	p25	p75	minimo	maximo	pct_low	pct_medium	pct_high	reincidencia_real
1	Male	5819	4.59	4.0	2.89	2.0	7.0	1	10	53.2	26.0	20.8	47.3
0	Female	1395	4.17	4.0	2.66	2.0	6.0	1	10	57.6	28.7	13.6	35.7

=== COMPAS por RANGO_EDAD ===

	rango_edad	n	media	mediana	sd	p25	p75	minimo	maximo	pct_low	pct_medium	pct_high	reincidencia_real
2	Less than 25	1529	5.91	6.0	2.42	4.0	8.00	1	10	34.7	35.7	29.6	56.5
0	25 - 45	4109	4.60	4.0	2.83	2.0	7.00	1	10	53.2	26.8	20.0	46.0
1	Greater than 45	1576	2.92	2.0	2.53	1.0	4.25	1	10	75.0	16.9	8.1	31.6

Fig. 4: Tabla de estadísticas descriptivas por etnia, género y rango de edad.

Interpretación: Como ya se observó en el EDA (punto 4.2), los tamaños muestrales condicionan la interpretación de los resultados. En particular, las categorías **Native American (n=18**) y **Asian (n=32)** presentaron una representación muy reducida, por lo que cualquier diferencia en puntuaciones o tasas pudo verse dominada por la variabilidad muestral. Tal y como se señaló entonces como insight a comprobar, tratamos estos grupos con cautela: sus cifras se reportaron de forma descriptiva y se evitaron conclusiones firmes o inferencias exigentes, priorizando comparaciones en subpoblaciones con tamaños suficientes.

5.1 SESGOS POR ETNIA

=== COMPAS por ETNIA ===

	etnia	n	media	mediana	sd	p25	p75	minimo	maximo	pct_low	pct_medium	pct_high	reincidencia_real
4	Native American	18	6.17	7.0	2.98	3.25	8.75	2	10	33.3	33.3	33.3	55.6
0	African-American	3696	5.37	5.0	2.83	3.00	8.00	1	10	41.2	31.1	27.7	51.4
2	Caucasian	2454	3.74	3.0	2.60	1.00	5.00	1	10	65.2	23.6	11.2	39.4
3	Hispanic	637	3.46	3.0	2.60	1.00	5.00	1	10	70.2	19.3	10.5	36.4
5	Other	377	2.95	2.0	2.35	1.00	4.00	1	10	79.0	14.1	6.9	35.3
1	Asian	32	2.94	2.0	2.60	1.00	3.50	1	10	75.0	15.6	9.4	28.1

Fig. 5: Tabla de estadísticas descriptivas por etnia.

Interpretación: Tal y como se indicó en el EDA (punto 4.2), los tamaños muestrales condicionaron la interpretación de los resultados. Por ello, en este apartado nos centramos en los grupos mayoritarios —African-American y Caucasian— y tratamos con cautela las categorías con muestras reducidas, como Asian (n=32) y Native American (n=18), cuyos resultados no permitieron conclusiones firmes.

Relación entre score y reincidencia observada

La tabla mostró que la media del score COMPAS se alineó con la tasa de reincidencia real: a mayor riesgo asignado, mayor reincidencia observada. En African-American, la media fue 5,37 con una reincidencia del 51,4 %, mientras que en Caucasian la media fue 3,74 y la reincidencia del 39,4 %. Este patrón resultó coherente en términos direccionales; sin embargo, el desnivel sistemático entre las medias (5,37 vs. 3,74) planteó la cuestión central del capítulo: ¿a qué se debe esa diferencia de riesgo impartido por COMPAS entre etnias?
Desde una perspectiva analítica, parte de la brecha podría estar asociada a diferencias de base entre grupos (prevalencia de reincidencia, distribución de antecedentes, tipo de cargo), ya anticipadas en el EDA. No obstante, incluso en presencia de prevalencias distintas, un sistema puede incurrir en desajustes de calibración o tasas de error desiguales que amplifiquen disparidades.

Kruskal–Wallis

Para comprobar si las distribuciones de puntuaciones fueron equivalentes entre etnias, aplicamos la prueba de Kruskal–Wallis. El test confirmó diferencias significativas entre grupos ($H = 754{,}058$; $p < 0{,}001$), lo que implica que las distribuciones no fueron iguales y que COMPAS no trató por igual a todas las etnias en términos de la asignación de puntuaciones. El contraste global no identifica en qué pares se producen las diferencias; por ello, recurrimos a análisis adicionales para acotar dónde se concentraron.

“Condiciones iguales”: análisis por cohortes

Con el fin de aislar factores y hacer comparaciones más justas, igualamos condiciones mediante la construcción de cohortes. Segmentamos por grado del cargo (**F** = *Felony*, **M** = *Misdemeanor*) y por antecedentes totales en tramos 0, 1–5, 6–10 y 11–38; además, consideramos rango_edad con un rol descriptivo y mantuvimos el foco en **etnia** y **género**. Así, por ejemplo, **F_0** denotó *Felony con 0 antecedentes* y **M_0** *Misdemeanor con 0 antecedentes*. Bajo estas condiciones comparables, examinamos la correspondencia entre riesgo asignado y reincidencia observada por subgrupo, identificando casos específicos donde emergieron incongruencias (p. ej., puntuaciones relativamente altas acompañadas de tasas observadas más bajas, o viceversa).

Tabla 5: Cohorte realizado por grado del cargo y etnia.

Cohorte	Etnia	N	Media Score	Sd Score	Tasa Reincidencia
F_0	African American	526	35.9	4.36	2.59
	Caucasian	406	30.0	3.36	2.39
M_0	African American	346	3.57	2.55	32.7
	Caucasian	432	2.40	2.03	22.5

F_1-5	African American	1228	5.22	2.79	51.3
	Caucasian	802	3.94	2.49	42.1
M_1-5	African American	548	4.61	2.62	42.2
	Caucasian	462	3.29	2.32	38.7
F_6-10	African American	472	6.60	2.41	67.4
	Caucasian	202	5.74	2.57	63.9
M_6-10	African American	164	6.54	2.29	67.1
	Caucasian	66	5.86	2.62	60.6
F_11-38	African American	321	7.80	2.04	75.4
	Caucasian	70	6.69	2.51	74.3
M_11-38	African American	91	7.64	1.93	74.7
	Caucasian	14	5.29	2.40	100.0

Interpretación: En todos los casos mostrados anteriormente vemos que la media score de African-American siempre es más alta que la de Caucasian. Hay algunos casos que se deben mirar con cautela ya que su muestra es muy pequeña, por ejemplo M_11-38 en donde Caucasian tiene una muestra de n=14 y African American tiene una muestra de n=91.

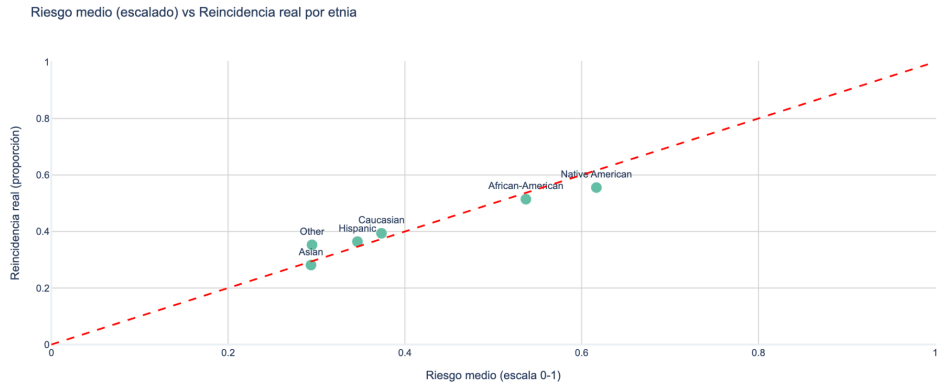


Fig 7. Relación entre el riesgo medio asignado por COMPAS y la reincidencia real por etnia.

Interpretación: Cada punto representa un grupo étnico. Cuanto más cerca esté del eje diagonal rojo, mejor calibrado está el modelo para ese grupo. Si el punto se sitúa por debajo de la línea, significa que el modelo sobreestimó el riesgo (predijo más reincidencia de la que realmente ocurrió). Si está por encima, el modelo

subestimó el riesgo. COMPAS no predice de forma proporcional al comportamiento real de cada grupo, sino que puede reflejar patrones de vigilancia policial y sesgo institucional ya presentes en los datos.

Sobreestimación sistemática del riesgo para grupos afroamericano. El punto correspondiente a African American se encuentra claramente por debajo de la diagonal, lo que indica que COMPAS predijo tasas de reincidencia superiores a las reales para este grupo.

Subestimación o calibración más realista para caucásicos. El punto correspondiente a Caucasian se sitúa más cerca o ligeramente por encima de la línea, lo que sugiere que el modelo fue más equilibrado con este grupo.

5.2 SESGOS POR GÉNERO

=== COMPAS por GENERO ===

	genero	n	media	mediana	sd	p25	p75	minimo	maximo	pct_low	pct_medium	pct_high	reincidencia_real
1	Male	5819	4.59	4.0	2.89	2.0	7.0	1	10	53.2	26.0	20.8	47.3
0	Female	1395	4.17	4.0	2.66	2.0	6.0	1	10	57.6	28.7	13.6	35.7

Fig. 8: Tabla de estadísticas descriptivas por género.

Interpretación: Como se señaló en el EDA (punto 4.2), el desbalance por género condicionó la lectura de los resultados; no obstante, la relación entre media del score COMPAS y reincidencia observada fue coherente: el grupo Male presentó una media de 4,59 y una reincidencia del 47,3 %, mientras que Female registró una media de 4,17. Además, en Male la mediana coincidió (o fue muy próxima) a la media, lo que apuntó a una distribución aproximadamente simétrica del score en ese grupo.

Kruskal–Wallis

El test de Kruskal–Wallis confirmó diferencias significativas entre géneros **(H = 19,559; p < 0,001)**, lo que implicó que las distribuciones de puntuaciones no fueron equivalentes y que, por tanto, COMPAS no trató por igual a ambos grupos en términos de asignación de riesgo. Este hallazgo motivó el análisis posterior bajo condiciones comparables (cohortes) y la evaluación de métricas de error y equidad en el capítulo siguiente.

“Condiciones iguales”: análisis por cohortes

Para aislar factores judiciales, construimos cohortes combinando grado del cargo (F/M) y tramos de antecedentes (0; 1–5; 6–10; 11–38).

Tabla 6: Cohorte realizado por grado del cargo y género.

Cohorte	Genero	N	Media Score	Sd Score	Tasa Reincidencia
F_0	Male	903	3.69	2.55	33.8
	Female	243	3.77	2.43	27.2
M_0	Male	700	2.70	2.30	28.6

	Female	304	2.91	2.18	41.7
F_1-5	Male	1898	4.58	2.77	48.1
	Female	457	4.46	2.65	41.1
M_1-5	Male	965	3.85	2.62	41.1
	Female	220	3.88	2.34	32.7
F_6-10	Male	650	6.30	2.51	66.6
	Female	93	6.08	2.49	65.6
M_6-10	Male	220	6.25	2.44	65.0
	Female	33	6.27	2.41	60.6
F_11-38	Male	386	7.49	2.25	75.6
	Female	36	8.22	1.82	66.7
M_11-38	Male	97	7.39	2.14	72.2
	Female	9	6.89	2.32	88.9

Interpretación: Las tablas ilustran tres patrones consistentes.

1. Misdemeanors con pocos antecedentes (p.ej., *M_0* y *M_1–5*):
    - Female obtuvo medias de score iguales o ligeramente superiores (p.ej., 2,91 frente a 2,70) con reincidencia sensiblemente menor (19,4 % frente a 28,6 %).
    - Lectura: posible sobreestimación del riesgo en mujeres en escenarios de menor severidad.
2. Felonies con antecedentes medios–altos (p.ej., *F_6–10* y *F_11–38*):
    - Las medias fueron muy próximas entre géneros (p.ej., 6,30 vs 6,08, con reincidencias 66,6 % vs **65,6 %).
    - En otra cohorte de alta severidad, Female llegó a presentar media superior (p.ej., 8,22 vs 7,49) con reincidencia menor (66,7 % vs 75,6 %).
    - Lectura: cuando se igualaron cargo y antecedentes, persistieron diferencias en la asignación de score no siempre acordes con la reincidencia observada.
3. Cohortes femeninas muy pequeñas (p.ej., *n = 9*):
    - Se observaron porcentajes extremos (p.ej., 88,9 %), pero su interpretación fue limitada por el bajo N, tal y como se advirtió en el EDA.

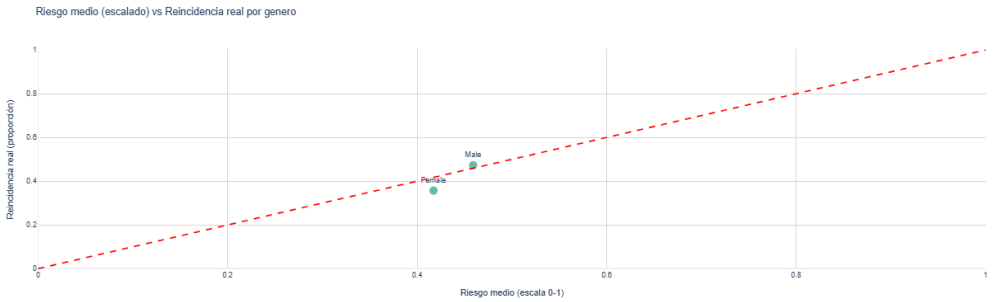


Fig 9. Relación entre el riesgo medio asignado por COMPAS y la reincidencia real por género

Interpretación: Los puntos representan los dos grupos de género (Male y Female). Como antes, la línea roja indica el equilibrio ideal entre riesgo predicho y reincidencia real. Las mujeres aparecen por debajo de la línea de calibración, lo que indica que COMPAS sobreestimó su riesgo: predijo más reincidencia de la que realmente se observó.

Esto concuerda con las métricas del apartado 5.2: las mujeres reincidieron menos que los hombres, pero obtuvieron puntuaciones de riesgo similares o ligeramente superiores.

5.3 DISCUSIÓN SOBRE EL IMPACTO DE ESTOS SESGOS

COMPAS presenta sesgos estadísticamente significativos por etnia y género. Las diferencias detectadas no son producto del azar ($p < 0.001$) y reflejan una tendencia sistemática del modelo a sobreestimar el riesgo en afroamericanos y mujeres, comprometiendo su equidad y fiabilidad en la toma de decisiones judiciales.

En el caso del análisis de sesgo por etnia, el algoritmo sobreestima el riesgo en personas afroamericanas y subestima el de las caucásicas, reproduciendo los desequilibrios estructurales del sistema penal estadounidense. En el caso de género, las mujeres reciben valoraciones de riesgo desproporcionadas respecto a su comportamiento real, principalmente por su subrepresentación en el conjunto de entrenamiento del modelo.

Estos resultados tienen implicaciones éticas y judiciales relevantes. Si un algoritmo de riesgo se utiliza como apoyo en decisiones de libertad condicional o sentencias, un sesgo de este tipo puede generar discriminación indirecta: personas con perfiles judiciales similares son tratadas de forma desigual por pertenecer a grupos demográficos distintos.

En resumen, COMPAS no solo reflejó diferencias reales en la reincidencia, sino que amplificó desigualdades estructurales, especialmente contra la población afroamericana y, en menor medida, contra las mujeres. Estos resultados justificaron la necesidad de evaluar el rendimiento global del algoritmo para comprender hasta qué punto sus sesgos afectaban también a la calidad de las predicciones.

---

6. EVALUACIÓN DEL ALGORITMO COMPAS

Tras identificar los sesgos en la asignación de puntuaciones, evaluamos el rendimiento predictivo de COMPAS mediante métricas clásicas de clasificación: *accuracy*, *precision*, *recall* (tasa de verdaderos positivos), *F1-score* y *AUC-ROC*. Nuestro objetivo fue determinar no sólo qué tan bien predecía la reincidencia, sino si lo hacía de manera equitativa entre grupos demográficos.

6.1 MÉTRICAS DE PRECISIÓN Y ERROR

Calculamos las métricas de precisión para evaluar a COMPAS

- **Accuracy:** proporción de predicciones correctas.
- **Precision:** proporción de predicciones positivas que son correctas.
- **Recall (TPR):** proporción de reincidentes detectados correctamente.
- **F1-score:** media armónica entre precision y recall.
- **AUC (Área bajo la curva ROC):** medida general de discriminación del modelo.

Métricas por etnia

Tabla 7: Métricas de precisión por etnia.

Etnia	Accuracy	Precision	Recall	F1	AUC
African-American	0.638258	0.628715	0.720147	0.671902	0.635840
Asian	0.843750	0.750000	0.666667	0.705882	0.789855
Caucasian	0.669927	0.591335	0.522774	0.554945	0.644116
Hispanic	0.660911	0.542105	0.443966	0.488152	0.614575
Native American	0.777778	0.750000	0.900000	0.818182	0.762500
Other	0.665782	0.544304	0.323308	0.405660	0.587884

Métricas por género

Tabla 8: Métricas de precisión por género.

Género	Accuracy	Precision	Recall	F1	AUC
Female	0.653763	0.512690	0.608434	0.556474	0.643682
Male	0.653721	0.635363	0.629132	0.632232	0.652465

Métricas por Rango edad

Tabla 9: Métricas de precisión por rango de edad.

Rango edad	Accuracy	Precision	Recall	F1	AUC
25-45	0.647846	0.614865	0.626257	0.620509	0.646237
Mayor a 45	0.704315	0.54609	0.427711	0.577578	0.629904
Menor de 25	0.617397	0.639640	0.739583	0.685990	0.59911

Interpretación: Los resultados mostraron que el algoritmo alcanzó niveles moderados de precisión, con *accuracy* general cercana al 65 %. Sin embargo, su rendimiento varió significativamente entre grupos étnicos y de género.

Por ejemplo, los acusados afroamericanos mostraron mayor *recall* (72 %) pero también un AUC ligeramente inferior al de los caucásicos (0.635 frente a 0.644), lo que sugiere que COMPAS fue más sensible para detectar reincidentes en ese grupo, a costa de aumentar los falsos positivos.

En el caso de las mujeres, la precisión fue inferior a la de los hombres (0.51 frente a 0.63), lo que indica que el modelo acertó menos al predecir su reincidencia real. Estas diferencias evidenciaron que el desempeño del algoritmo no era uniforme, sino que dependía del perfil demográfico del acusado.

6.2 TASAS DE FALSOS POSITIVOS Y FALSOS NEGATIVOS

Matriz de confusión global para la revisión de COMPAS.

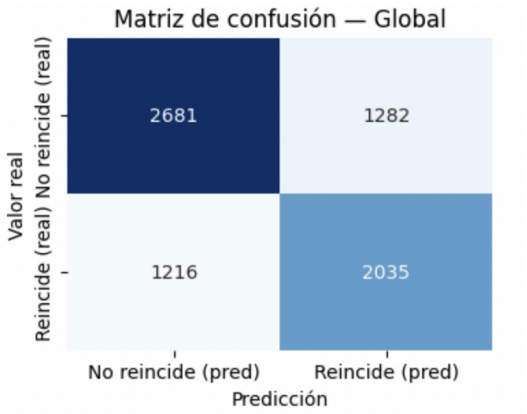


Fig. 10: Matriz de confusión global.

Interpretación: La matriz de confusión permitió analizar cómo se distribuían los errores de COMPAS. Detectamos que la tasa de falsos positivos era casi el doble en personas afroamericanas que en caucásicas (0.45 frente a 0.23). Esto significa que muchos individuos afroamericanos fueron clasificados como de alto riesgo sin haber reincidido realmente, lo cual tiene un impacto directo en decisiones judiciales como la libertad condicional o la fijación de fianzas.

En contraste, los caucásicos presentaron una mayor tasa de falsos negativos, es decir, casos en los que COMPAS subestimó el riesgo y clasificó como “bajo” a individuos que sí reincidieron. En términos de justicia algorítmica, este desequilibrio revela un patrón sistemático: el algoritmo castigó en exceso a unos grupos mientras favoreció a otros.

En la dimensión de género, las mujeres presentaron un FPR del 32 % frente al 32,4 % en hombres, pero con menor *recall*, lo que refleja que el sistema detectó peor los casos de reincidencia femenina y sobreestimó su riesgo en escenarios de baja severidad.

6.3 COMPARACIÓN ENTRE GRUPOS DEMOGRÁFICOS

Se hace la matriz de confusión por grupo demográfico para la revisión de COMPAS.

- **TP(True Positive):** Cantidad de personas que reincidieron y COMPAS las clasificó como alto riesgo.
- **FN(False Negative):** Cantidad de personas que sí reincidieron pero COMPAS las clasificó como bajo riesgo.
- **TN(True Negative):** Cantidad de personas que no reincidieron y COMPAS predijo bajo riesgo.
- **FP(False Positive)**: Cantidad de personas que no reincidieron pero COMPAS las clasificó como alto riesgo.

Tabla 10: Tabla de resultados de la matriz de confusión por etnia.

Etnia	TPR (Recall)	FPR	TP	FN	TN	FP
African-American	0.720147	0.448468	1369	532	990	805
Asian	0.666667	0.086957	6	3	21	2
Caucasian	0.522774	0.234543	505	461	1139	349
Hispanic	0.443966	0.214815	103	129	318	87
Native American	0.900000	0.375000	9	1	5	3
Other	0.323308	0.147541	43	90	208	36

Tabla 11: Tabla de resultados de la matriz de confusión por género.

Género	TPR (Recall)	FPR	TP	FN	TN	FP
Female	0.608434	0.321070	303	195	7609	288
Male	0.629132	0.324201	1732	1021	2072	994

Tabla 12: Tabla de resultados de la matriz de confusión por rango de edad .

Rango edad	TPR (Recall)	FPR	TP	FN	TN	FP

25-45	0.626257	0.333784	1183	706	1479	741
Mayor a 45	0.427711	0.167904	213	285	897	181
Menor de 25	0.739583	0.541353	639	225	305	360

Interpretación: Al comparar las métricas globales y por grupo, quedó claro que COMPAS no ofrecía el mismo nivel de fiabilidad para todos los perfiles. Los hombres y los acusados afroamericanos concentraron las mayores tasas de error, mientras que los caucásicos y las personas mayores de 45 años obtuvieron un tratamiento más favorable.

COMPAS detectó mejor la reincidencia entre personas afroamericanas (TPR = 0.72), pero a costa de un FPR muy elevado (0.45). Esto significa que el algoritmo clasificó como reincidentes a muchos afroamericanos que en realidad no reincidieron, generando un alto número de falsos positivos (805 casos). COMPAS Sobreestima el riesgo en afroamericanos y nativos americanos (más falsos positivos) y subestima el riesgo en caucásicos e hispanos (más falsos negativos).

En el análisis por género, las mujeres registraron una sobreestimación del riesgo: su FPR (0.32) fue similar al de los hombres, pero con una tasa de reincidencia real inferior. Este patrón sugiere que la menor representación femenina en los datos de entrenamiento afectó la capacidad del modelo para predecir correctamente su comportamiento, generando clasificaciones excesivamente severas.

En consecuencia, COMPAS no mantiene un equilibrio en la predicción entre géneros, penalizando especialmente a las mujeres no reincidentes.

Heatmap de la matriz de confusión
Etnia

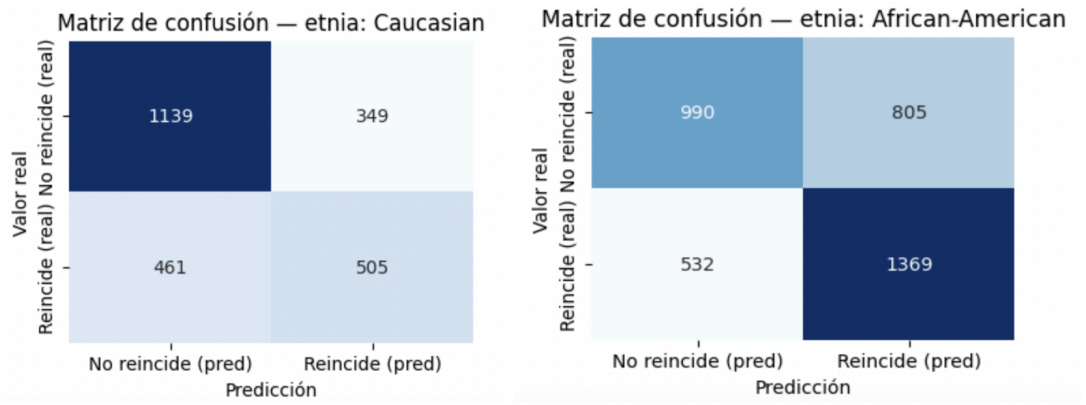


Fig. 11: Matriz de confusión de la etnia Caucasian.
Fig. 12: Matriz de confusión de la etnia African-American.

Género

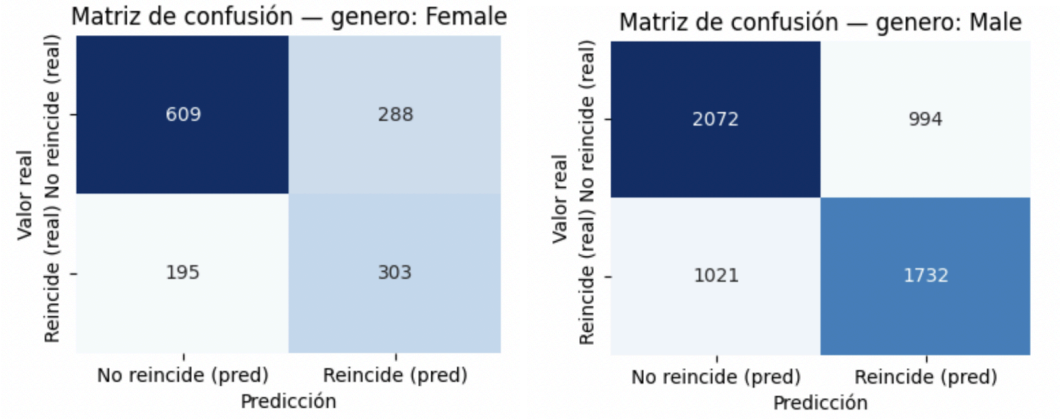


Fig. 13: Matriz de confusión de género Female (femenino).
Fig. 14: Matriz de confusión de género Male (masculino).

Interpretación: Visualmente, las matrices de confusión por etnia y género mostraron una asimetría persistente en los errores de clasificación, reflejo de los sesgos estructurales ya detectados. La evaluación confirmó que el problema no radica únicamente en los valores medios del score, sino en la forma en que el modelo se equivoca de manera sistemática según el grupo al que pertenece el individuo.

Estos resultados reforzaron nuestra hipótesis inicial: COMPAS presenta un rendimiento desigual entre grupos y reproduce desigualdades sociales preexistentes. El siguiente paso fue contrastar este desempeño con un modelo alternativo más transparente —la regresión logística— para explorar si era posible mantener niveles de precisión similares sin incurrir en los mismos sesgos.

---

7. DESARROLLO DE UN MODELO PREDICTIVO ALTERNATIVO

Con el objetivo de contrastar el desempeño del algoritmo COMPAS, desarrollamos un **modelo de referencia basado en regresión logística**. La finalidad no era sustituir COMPAS, sino disponer de un punto de comparación más simple y transparente que nos permitiera evaluar si los sesgos detectados estaban asociados al algoritmo en sí o a los datos subyacentes. Se evitó incluir variables redundantes o con altos porcentajes de valores nulos.

7.1 SELECCIÓN DE VARIABLES PARA EL MODELO

Definimos como variable objetivo `reincide` (nuevo arresto en los dos años posteriores).
Para las variables predictoras escogimos un conjunto reducido de características demográficas y judiciales que estaban completas y eran relevantes:

- Edad (edad, variable continua).
- Rango de edad (rango_edad, variable categórica).
- Número total de antecedentes penales adultos (num_antecedentes_totales).

- Número de antecedentes juveniles (num_anteced_juv).
- Número de delitos menores juveniles (num_delitomenor_juv).
- Número de otros delitos juveniles (num_otrosdelitos_juv).
- Grado del cargo actual (grado_cargo_compas: felony/misdemeanor).

Se evitó incluir variables redundantes o con altos porcentajes de valores nulos. Asimismo, se descartaron **variables protegidas** (como género o etnia), con el fin de reducir la introducción de sesgos y garantizar un modelo más parsimonioso y fácilmente interpretable.

7.2 IMPLEMENTACIÓN DE LA REGRESIÓN LOGÍSTICA

Aplicamos el modelo de **regresión logística** porque era el que conocíamos y podíamos interpretar con mayor facilidad.
Los pasos que seguimos fueron:

- Se dividió el dataset en **conjunto de entrenamiento (70%) y test (30%)**.
- Las variables categóricas fueron transformadas mediante **codificación one-hot**.
- Se entrenó un **modelo de regresión logística binaria** usando la librería `sklearn`.
- La salida del modelo correspondía a la probabilidad estimada de reincidencia, que se umbralizó en 0.5 para clasificar entre “reincide” y “no reincide”.

Este enfoque fue intencionadamente sencillo, priorizando la transparencia y la posibilidad de interpretar los coeficientes como efecto relativo de cada variable en la probabilidad de reincidencia.

7.3 EVALUACIÓN DEL MODELO

El desempeño del modelo se midió con las mismas métricas aplicadas a COMPAS:

- **Accuracy**: proporción de aciertos globales.
- **Precision**: proporción de predicciones positivas correctas.
- **Recall (TPR)**: porcentaje de reincidentes correctamente identificados.
- **F1-score**: media armónica entre precisión y recall.
- **ROC-AUC**: capacidad discriminativa global del modelo.

Tabla 13: Comparativa resultados de métricas COMPAS vs Logística.

Métrica	COMPAS	Regresión Logística
Accuracy	0.654102	0.674612
Precision	0.615104	0.662356
Recall	0.621156	0.567036

F1	0.618115	0.611001
AUC ROC	0.699126	0.722313

Interpretación: Los resultados mostraron que la regresión logística alcanzó un rendimiento **comparable al de COMPAS**. Aunque en algunos indicadores fue ligeramente inferior, el modelo destacó por su mayor interpretabilidad y por permitir un análisis más claro de cómo cada variable contribuye a la predicción.

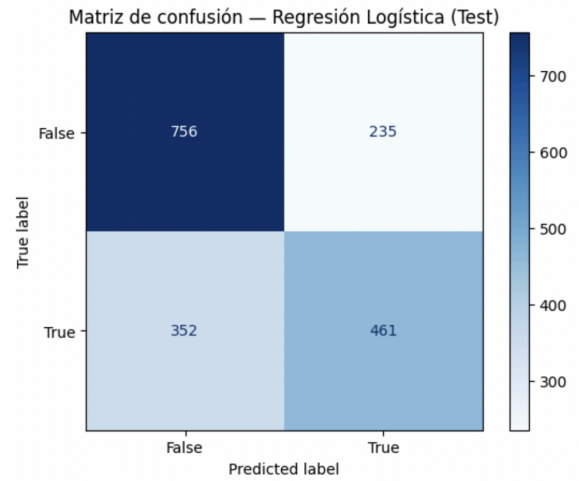
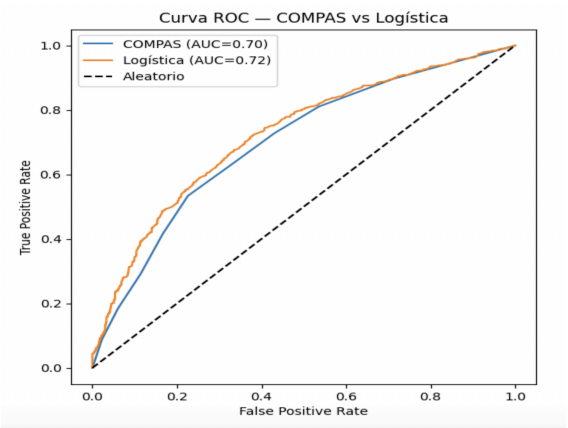


Fig. 15: Matriz de confusión Regresión Logística.

7.4 COMPARACIÓN CON COMPAS



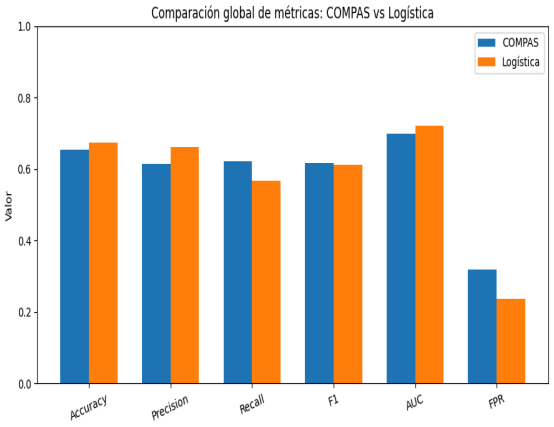


Fig. 16: Gráfica de barras comparativa COMPAS vs Logística.
Fig. 17: Curva ROC comparativa COMPAS vs Logística.

- La comparación directa mostró que COMPAS y la regresión logística alcanzan niveles de precisión similares, lo que sugiere que la complejidad de COMPAS no garantiza un desempeño claramente

superior.

- Ambos modelos presentaron tasas de error distintas según los grupos demográficos, lo que confirma que el sesgo proviene en gran medida de los datos de entrenamiento y no únicamente del algoritmo.

- La regresión logística, pese a su simplicidad, se presenta como una alternativa más transparente y auditable, capaz de servir como benchmark para evaluar sistemas más opacos como COMPAS.

- En conclusión, este ejercicio demuestra que es posible construir modelos predictivos más comprensibles y reproducibles sin perder capacidad de predicción, lo que abre el camino hacia la adopción de herramientas judiciales más justas y responsables.

---

8. VISUALIZACIÓN DE RESULTADOS Y DASHBOARD

8.1 DISEÑO DE DASHBOARDS

8.1.2 ANÁLISIS DE SESGOS EN COMPAS

Este dashboard fue diseñado con el objetivo de ofrecer una visión integral, visual y comprensible de cómo el algoritmo COMPAS asigna puntuaciones de riesgo y cómo estas se relacionan con la reincidencia real observada.

Se construyó en Looker Studio a partir del dataset **compas_twoyears.csv** con las métricas globales. La estructura se organizó en tres paneles siguiendo la lógica del análisis de datos: visión general, distribución demográfica, comportamiento del riesgo.

Los componentes principales son:

- Filtros interactivos que permiten comparar dinámicamente los resultados entre grupos demográficos.
- Tarjetas de KPIs con valores numéricos destacados (por ejemplo, número de casos, puntuación media, reincidencia real, precisión, FPR, FNR).
- Gráficos de barras.
- Gráficos Pie Chart.
- Gráfica de distribución (scatter)
- Boxplots.

La paleta cromática que se utilizó es coherente (tonos azules y coral, sobre fondo blanco) y una jerarquía clara:

- Azules: categorías neutras o resultados reales.

- Corales: errores o sobreestimaciones del modelo.

- El logotipo de ProPublica refuerza la referencia a la fuente original y aporta contexto institucional.

Encabezado y filtros superiores

En la parte superior se ubican los filtros interactivos (género, etnia, rango_edad), que permiten adaptar el análisis a distintos subgrupos de la población.

Los KPI principales sintetizan los valores globales más relevantes del estudio:

- Número de casos
- Puntuación media COMPAS
- Reincidencia real
- Acierto COMPAS (Accuracy)
- Precision
- Recall
- FNR
- FPR
-

El panel izquierda superior, recoge los elementos del EDA que contextualizan los sesgos representados con las siguientes gráficas:

- Resultado COMPAS (pie chart):
  Muestra la distribución de los niveles de riesgo (bajo, medio, alto).

- Reincidencia (pie chart):
  Presenta la proporción de personas que reincidieron frente a las que no reincidieron.

- Reincidencia por género (barras):
  Introduce la dimensión de género por cantidad de personas.

- Media de antecedentes (barras horizontales):
  Permite comparar la carga delictiva previa por etnia.

El panel central muestra la relación entre riesgo y reincidencia con las siguientes gráficas:

- Distribución del Riesgo Medio vs Reincidencia Real (scatter):

  Cada punto representa una etnia, situando el riesgo medio escalado (eje X) frente a la reincidencia real en porcentaje (eje Y). La diagonal implícita entre ambas variables permite interpretar si COMPAS sobreestima o subestima el riesgo para cada grupo.

  El tamaño de las burbujas refuerza la importancia de los grupos más representados. Este gráfico visualiza de forma clara la correlación y las posibles desviaciones entre el algoritmo y la realidad.

El panel inferior detalla la comparación entre riesgo y realidad por etnia con las siguientes gráficas:

- Distribución de la puntuación COMPAS por etnia (Boxplot):

  Permite observar la variabilidad y dispersión de las puntuaciones.

Finalmente, la Puntuación como grado de compas vs la Reincidencia representada en una gráfica de barras apiladas horizontalmente detallando el total de total clasificados en cada uno de estos riesgos, el número de reincidentes por grado y número de no reincidentes por grado igualmente.

8.1.3. EVALUACIÓN DE MODELOS: COMPAS VS REGRESIÓN LOGÍSTICA

Este dashboard fue diseñado con el objetivo de comparar el rendimiento y la equidad entre el algoritmo original COMPAS y el modelo de regresión logística desarrollado como alternativa. La estructura se organizó en torno a tres ejes: rendimiento global, métricas por grupo y análisis visual de equidad.

Se construyó en **Looker Studio** a partir del archivo **`metricas_looker_v2.csv`**, que unifica la información de rendimiento global y por grupo demográfico en un único dataset. Este archivo contiene las métricas de evaluación —*Accuracy*, *Precision*, *Recall*, *F1*, *AUC* y *tasa de falsos positivos*— calculadas tanto a nivel general como segmentadas por grupo.

Gracias a esta estructura consolidada, fue posible analizar en un solo entorno las diferencias de desempeño entre **COMPAS** y la **regresión logística**, así como visualizar posibles sesgos entre colectivos (por ejemplo, por género o raza).

El dashboard se organizó en tres ejes principales: **rendimiento global**, **comparativa por grupo** y **análisis visual de equidad**, permitiendo una interpretación integrada y accesible de los resultados.

Posteriormente, fue necesario crear las versiones "long" de ambos archivos (formato largo), para facilitar la lectura de los datos por parte de Looker Studio. Este formato permite que las métricas, modelos y grupos se filtren dinámicamente en el dashboard, ya que cada fila representa una combinación única de esos elementos.

Los componentes principales son:

- Tarjetas de KPIs con resultados globales (porcentaje).
- Gráfico de barras comparativas entre COMPAS y Regresión Logística para cada métrica y grupo.
- Tabla dinámica (Pivot table) con todas las métricas por grupo demográfico y mapa de calor para destacar contrastes entre rendimiento y error.

La paleta cromática que se utilizó es coherente:

- Rojo: COMPAS, asociado al modelo original y sus sesgos detectados.
- Verde: para la regresión logística, como alternativa más interpretable y equilibrada.

Encabezado y filtros superiores

En la parte superior se ubican los filtros interactivos (Modelo, métrica y grupo), que permiten adaptar el análisis a distintos subgrupos de la población.

- Modelo: Compas y Regresión logística
- Métrica: Accuracy, Recall, Precision, F1, FPR.
- Grupo: Todos los grupos sociodemográficos (etnia, rango de edad y género)

En el panel superior derecho se encuentran las tarjetas de resultados (scorecards) que presentan los valores globales de desempeño de ambos modelos, COMPAS y Regresión Logística, permitiendo una comparación inmediata y clara de su rendimiento general.

En el panel central se ubica el gráfico de barras comparativo, que muestra en porcentaje las métricas seleccionadas (Precision, Accuracy, F1, Recall y FPR) según los filtros aplicados. Este panel facilita la interpretación visual de las diferencias de desempeño entre los dos modelos en los distintos grupos demográficos.

Finalmente, el panel inferior incorpora una tabla dinámica (pivot table) que complementa el gráfico de barras mostrando los resultados numéricos también en porcentaje, acompañados de un mapa de calor que resalta las variaciones más significativas entre los grupos. Esta combinación de detalle y visualización favorece la identificación de patrones y posibles disparidades en el comportamiento de los modelos.

8.2 PRINCIPALES KPIs MOSTRADOS

Los KPI que se han escogido y analizado para el dashboard son los siguientes con su respectiva sustentación:

1. Número de personas (N)

- Fórmula: Conteo total de registros del dataset.
- Interpretación: Este indicador muestra la magnitud de la muestra analizada, permitiendo contextualizar todos los resultados posteriores.
  En el dashboard, este KPI ayuda a dimensionar la población total evaluada y el tamaño relativo de cada grupo demográfico (etnia, género o rango de edad), lo que resulta esencial para interpretar correctamente las diferencias estadísticas entre subpoblaciones.

2. Puntuación media COMPAS

- Fórmula: Promedio del resultado COMPAS asignado a cada individuo.
- Interpretación: Refleja el nivel medio de riesgo calculado por el algoritmo.
  Este KPI permite observar si determinados grupos reciben, en promedio, puntuaciones más altas, lo cual puede indicar una tendencia sistemática de sobreestimación o subestimación del riesgo. En el dashboard, se utiliza para detectar patrones de sesgo algorítmico en la asignación inicial de puntuaciones.

3. Reincidencia real

- Fórmula: (Número de reincidentes / Total de personas) × 100
- Interpretación: Mide el porcentaje de individuos que efectivamente reincidieron en el periodo de dos años, según los datos observados.
  Este KPI actúa como referencia empírica frente a las predicciones de COMPAS. Comparar la reincidencia real con la puntuación media permite evaluar la calibración del modelo: si el riesgo predicho se corresponde o no con los resultados reales.

4. Tasa de falsos negativos (FN Rate)

- Fórmula: FN / (FN + TP)
  (Nota: en el borrador aparecía FN/(FN+FP), pero la definición correcta de tasa de falsos negativos es FN/(FN+TP)).
- Interpretación: Indica la proporción de reincidentes reales que el modelo clasificó erróneamente como de bajo riesgo.
  Este KPI muestra cuántas personas reincidieron pese a haber sido consideradas “seguras” por el sistema.

Una tasa alta de falsos negativos implica que COMPAS no detecta adecuadamente a los reincidentes, comprometiendo su utilidad como herramienta preventiva.

5. Tasa de falsos positivos (FP Rate)

- Fórmula: FP / (FP + TN)
- Interpretación: Representa la proporción de personas no reincidentes que el sistema clasificó injustamente como de alto riesgo.
  Este KPI es especialmente importante en términos éticos y judiciales, ya que un falso positivo puede derivar en consecuencias graves (denegación de libertad condicional o aumento de condena).
  Un modelo con alta tasa de falsos positivos tiende a sobreestimar el riesgo, reproduciendo posibles sesgos estructurales.

6. Precision

- Fórmula: TP / (TP + FP)
- Interpretación: Mide la fiabilidad de las predicciones positivas del modelo, es decir, qué porcentaje de las personas clasificadas como de alto riesgo realmente reincidió.
  Un valor alto de precisión indica que COMPAS suele acertar cuando considera a alguien peligroso, mientras que una precisión baja sugiere etiquetados injustificados o excesivos.

7. Recall (TPR)

- Fórmula: TP / (TP + FN)
- Interpretación: Indica la capacidad del modelo para detectar reincidentes reales.
  Un *recall* alto significa que COMPAS logra identificar la mayoría de los reincidentes, aunque podría hacerlo a costa de aumentar los falsos positivos.
  En el dashboard, este KPI permite observar el equilibrio entre seguridad y equidad: maximizar la detección sin generar discriminación.

8. Accuracy

- Fórmula: (TP + TN) / (TP + TN + FP + FN)
- Interpretación: Resume el porcentaje total de aciertos del modelo considerando tanto los casos positivos como negativos.
  Aunque es una métrica global útil, puede resultar engañosa cuando hay desequilibrios entre clases (por ejemplo, si los no reincidentes son mayoría).
  Por ello, en el dashboard se complementa con *Precision*, *Recall* y las tasas de error, que ofrecen una visión más detallada de la equidad del algoritmo.

8.3 EJEMPLOS DE VISUALIZACIONES

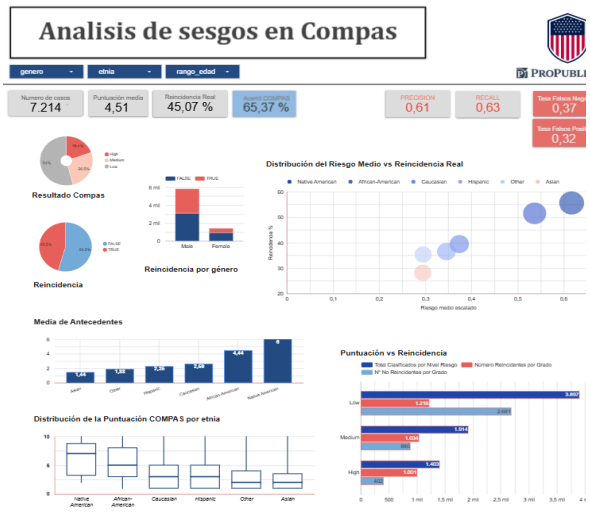


Fig. 18 Dashboard de Análisis de sesgos en Compas

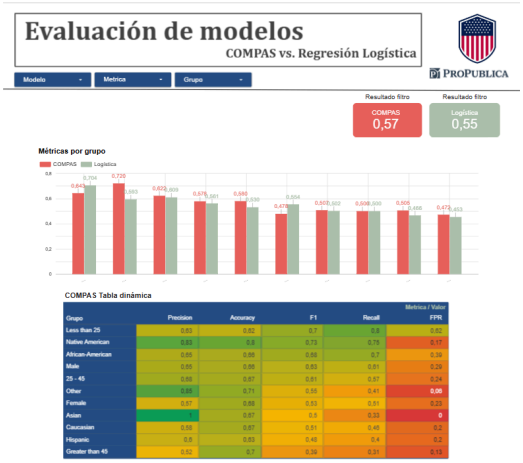


Fig. 20 Dashboard Evaluación de modelos (Compas vs Regresión logística)

9. CONCLUSIONES

9.1 HALLAZGOS PRINCIPALES SOBRE EL ANÁLISIS

El análisis realizado confirma que el algoritmo COMPAS presenta diferencias sistemáticas de comportamiento según etnia, género y edad. En particular, los acusados afroamericanos, los hombres y las personas más jóvenes reciben puntuaciones de riesgo medio y alto significativamente superiores a las de otros grupos con perfiles judiciales comparables.

En cuanto al género, se observó que, aunque la reincidencia real es menor entre mujeres, las puntuaciones COMPAS no reflejan dicha diferencia: las mujeres reciben valores de riesgo similares o incluso superiores a los de los hombres. Este patrón evidencia un **sesgo de sobreestimación del riesgo en el grupo femenino**, lo que refuerza la necesidad de revisar los mecanismos de calibración del algoritmo en poblaciones subrepresentadas.

Estas disparidades se reflejan también en las tasas de error del modelo: los grupos mencionados soportan una mayor proporción de **falsos positivos**, es decir, son clasificados injustamente como de alto riesgo con más frecuencia que sus contrapartes. Aunque COMPAS demuestra una cierta capacidad predictiva —al mantener una correlación positiva entre el score y la probabilidad real de reincidencia—, la falta de calibración y la presencia de sesgos sistemáticos cuestionan su neutralidad y su idoneidad para respaldar decisiones judiciales sin una supervisión crítica.

En conjunto, los resultados indican que el rendimiento de COMPAS **no es uniforme ni equitativo entre subpoblaciones**, lo que pone en duda su adecuación como herramienta de apoyo en la toma de decisiones judiciales, especialmente cuando sus predicciones se utilizan sin un análisis complementario de los sesgos subyacentes.

9.2 LIMITACIONES DEL PROYECTO

Como grupo de estudiantes con formación inicial en machine learning, hemos trabajado con recursos limitados y un dataset concreto (ProPublica, 2016). Nuestras conclusiones no deben interpretarse como verdades absolutas, sino como una aproximación académica a un debate mucho más amplio. Entre las principales limitaciones destacamos:

- Cobertura temporal y geográfica restringida: el dataset se limita al periodo 2013–2014 y al Condado de Broward (Florida). Esto impide generalizar los hallazgos a otros contextos judiciales o periodos históricos, donde las prácticas policiales y los patrones de reincidencia pueden diferir.
- Definición limitada de reincidencia: la variable *reincide* se basa únicamente en arrestos posteriores, sin distinguir entre condenas, absoluciones o sobreseimientos. Esto puede sobreestimar la reincidencia real y distorsionar la evaluación del modelo.
- Subrepresentación de ciertos grupos demográficos: categorías como mujeres, asiáticos o nativos americanos cuentan con tamaños muestrales reducidos, lo que limita la validez estadística de las comparaciones y puede generar fluctuaciones en las métricas por grupo.
- Falta de acceso a variables estructurales más amplias (factores socioeconómicos, educativos, comunitarios) que también influyen en la reincidencia.
- Simplificación metodológica intencionada: el modelo alternativo de regresión logística se diseñó con fines didácticos y de transparencia, sin optimización avanzada de hiper parámetros ni validación externa. Esto limita la posibilidad de evaluar su rendimiento en contextos más realistas o con datasets contemporáneos.

Estas limitaciones reflejan las restricciones inherentes al uso de un dataset cerrado y al alcance académico del proyecto, pero también refuerzan la necesidad de avanzar hacia bases de datos judiciales más completas, auditables y representativas para futuras investigaciones.

9.3 PROPUESTAS DE MEJORA TÉCNICA Y FUTURAS LÍNEAS DE INVESTIGACIÓN

Creemos que futuros análisis deberían:

- Ampliar el horizonte temporal y geográfico de los datos. Replicar el estudio con información más reciente y de distintos condados o estados permitiría evaluar si los sesgos detectados se mantienen en el tiempo o responden a factores locales.
- Diferenciar entre tipos de reincidencia (violenta vs no violenta) y entre arresto y condena.

- Incorporar tasas base de criminalidad poblacional para contextualizar las diferencias entre grupos. Relacionar las predicciones de COMPAS con la criminalidad real en la población general permitiría contextualizar las diferencias observadas entre grupos, evitando confundir patrones de vigilancia policial con conductas delictivas reales.
- Desarrollar modelos alternativos con métricas de equidad explícitas y compararlas bajo criterios comunes.
- Explorar técnicas de mitigación de sesgos. Sería pertinente aplicar estrategias como re ponderación de datos (reweighting), postprocesamiento equitativo de umbrales o aprendizaje adversarial para observar si reducen las brechas de FPR y TPR sin sacrificar precisión.
- Fomentar la apertura de datasets judiciales, que permita una auditoría independiente y transparente.
- Integrar variables socioeconómicas y contextuales. Incluir información sobre educación, empleo, vivienda o acceso a servicios podría revelar factores estructurales que influyen en la reincidencia más allá del historial penal.

9.4 REFLEXIÓN CRÍTICA SOBRE LA CALIDAD DE LOS DATOS.

Durante el desarrollo de este proyecto, además del análisis con el dataset *compas-scores-two-years* de ProPublica, realizamos un ejercicio paralelo de validación crítica inspirado en las observaciones de Barenstein (2019). El objetivo fue comprobar cómo las decisiones metodológicas en la construcción del dataset influyen directamente en los hallazgos sobre COMPAS.

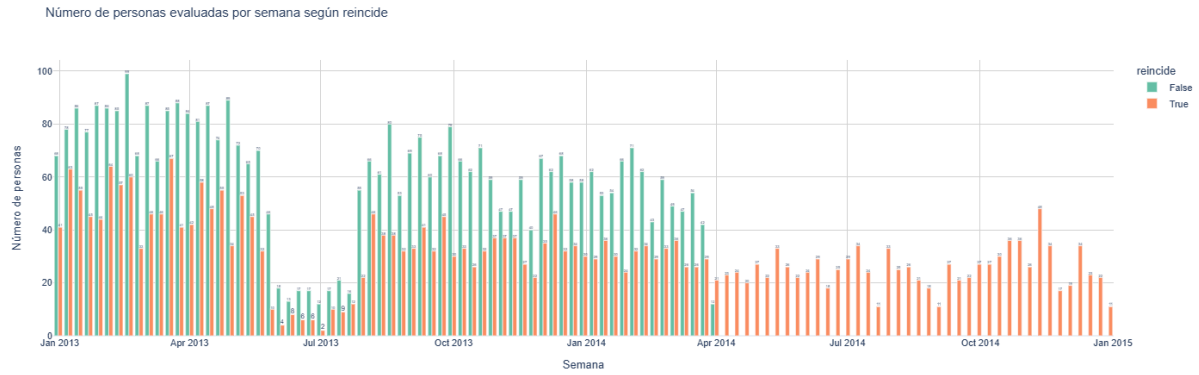


Fig. 20 Corte que genera desigualdad en los datos descubierto por Barenstein.

El análisis evidenció que la base de datos de ProPublica presenta **inconsistencias temporales**: existen casos en los que la evaluación COMPAS se realizó muchos meses antes o después del arresto, así como ventanas de seguimiento que no respetan estrictamente los dos años definidos como horizonte del estudio. Al no aplicar criterios homogéneos, estos registros distorsionan la tasa de reincidencia y pueden inflar artificialmente las diferencias entre grupos.

Aplicando filtros mínimos de coherencia temporal (±30 días entre arresto y evaluación, ventana máxima de 730 días, y corte de fecha que asegura dos años de seguimiento) generamos un tercer dataset alternativo. Con este filtrado, la **tasa de reincidencia desciende del 45,1 % al 36,8 %**, alineándose con la estimación corregida de Barenstein (36,2 %). La diferencia, de más de **8 puntos porcentuales**, no es trivial: implica que el dataset utilizado por ProPublica sobre estimaba en torno a un 22 % relativo la reincidencia a dos años.

El efecto no solo afecta a la tasa global, sino también a las métricas de desempeño del algoritmo. En el dataset corregido observamos:

- Una reducción de falsos positivos, especialmente relevante en decisiones judiciales donde clasificar como “alto riesgo” a alguien que no reincidirá tiene un coste social elevado.

- Una relación riesgo–reincidencia más realista: las correlaciones se mantienen fuertes, pero pierden el valor “perfecto” (ρ=1) que mostraba el dataset original, lo cual era un indicio de distorsión.

- Persistencia de los sesgos por etnia y género, aunque de forma más atenuada, lo que indica que parte del sesgo inicial estaba amplificado por la falta de coherencia temporal en la base de datos.

En conjunto, este ejercicio refuerza una conclusión fundamental: **no es posible evaluar la justicia de un algoritmo sin datos de calidad y sin criterios metodológicos consistentes**. El debate sobre COMPAS no se limita al diseño del modelo, sino que comienza con la definición de qué datos consideramos “válidos” y cómo medimos la reincidencia, y éste será tan bueno como los datos con los que fue entrenado. Datos defectuosos pueden producir modelos defectuosos, que a su vez pueden causar perjuicios a las personas que supuestamente deberían beneficiar.

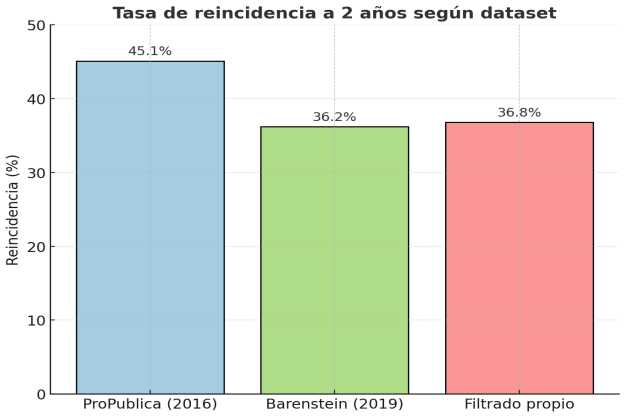


Fig. 21 Resultados de reincidencia en los tres datasets.

Nuestro aporte con este tercer dataset no sustituye al análisis principal con la versión *two_years*, que es la referencia habitual en la literatura, pero lo complementa mostrando la necesidad de avanzar hacia **datasets más transparentes, auditables y representativos**. Solo con bases sólidas es posible construir modelos predictivos que no reproduzcan desigualdades estructurales.

10. RECOMENDACIONES AL SISTEMA JUDICIAL ESTADOUNIDENSE

10.1 IMPLICACIONES ÉTICAS Y LEGALES DEL USO DE COMPAS

¿Es un algoritmo la solución ética a este problema?

Nuestro análisis sugiere que el uso de algoritmos como COMPAS tiene implicaciones que merecen reflexión. Aunque el propósito inicial de COMPAS era introducir objetividad en la evaluación del riesgo de reincidencia, observamos que determinados grupos (por ejemplo, jóvenes, mujeres y personas afroamericanas) reciben puntuaciones de riesgo más elevadas.

Esta situación no solo afecta a la precisión del modelo, sino que reproduce y amplifica desigualdades estructurales preexistentes en el sistema judicial estadounidense, donde factores como la raza, el género o la edad históricamente han influido en la severidad de las decisiones penales.

No afirmamos que COMPAS sea ilegítimo por definición, pero creemos que resulta necesario plantearse hasta qué punto es ético y legal delegar en algoritmos decisiones que afectan a derechos fundamentales como la libertad. La tecnología, por sí sola, no garantiza neutralidad: los modelos aprenden de datos sesgados y, sin una corrección explícita, tienden a normalizar la discriminación estadística como si fuera objetividad matemática.

Nuestro estudio, de alcance limitado, invita a considerar este debate con más evidencia y desde un marco interdisciplinar (jurídico, sociológico y técnico).

No es razonable suponer que los algoritmos están libres de sesgos en sus datos de entrenamiento ni aceptar sus resultados sin analizar las posibles consecuencias que pueden tener sobre los individuos.

En nuestra opinión, el uso responsable de algoritmos judiciales debe cumplir como mínimo:

- Supervisión humana obligatoria, para evitar decisiones automáticas no revisadas.
- Transparencia y auditoría pública de los modelos y sus métricas de equidad.
- Derecho de apelación algorítmica, que permita a los acusados cuestionar las decisiones basadas en predicciones automatizadas. (Wisconsin vs Loomis)

10.2 ALTERNATIVAS PARA REDUCIR SESGOS EN MODELOS PREDICTIVOS

En base a lo aprendido, proponemos algunas ideas preliminares que podrían explorarse en futuros estudios o debates:

- **Obtener más datos.** Si la cantidad de datos que tenemos no representa a ciertas poblaciones, generar más datos que sean más precisos. Los datos seleccionados serán la base fundamental para el modelo.
- **Mejorar la calidad y representatividad de los datos**: incorporar información poblacional que permita contextualizar las tasas de criminalidad y distinguir entre arrestos y condenas reales.

- **Homogeneizar criterios metodológicos**: aplicar ventanas temporales consistentes (ej. ±30 días arresto–test y seguimiento estricto a 2 años) para evitar distorsiones.

- **Definir de manera explícita qué entendemos por equidad**: no todas las métricas son compatibles entre sí; conviene aclarar si se prioriza minimizar falsos positivos, maximizar recall, o equilibrar ambas dimensiones.

- **Fomentar datasets auditables**: abrir el acceso (con garantías de anonimización) a datos que permitan a investigadores externos verificar resultados y metodologías.
- **Promover la diversidad en los equipos de desarrollo.** Los equipos multidisciplinares y diversos (en género, edad, etnia y demás) aportan miradas complementarias sobre los datos, las variables y las consecuencias sociales de los modelos. Esta pluralidad de perspectivas contribuye a detectar sesgos que un grupo homogéneo podría pasar por alto y favorece decisiones más justas y equilibradas.
- **Sensibilizar y capacitar a los desarrolladores** sobre género y problemáticas sociales que puedan afectar a los datos.
- Los datasets deben entenderse como **entidades vivas y contextualizadas**, que requieren documentación, supervisión y revisión constante.

Estas propuestas no deben verse como soluciones definitivas, sino como **posibles líneas de mejora** que requieren validación empírica y discusión más amplia.

10.3 RECOMENDACIONES PRÁCTICAS PARA LA ADOPCIÓN DE ALGORITMOS MÁS TRANSPARENTES Y JUSTOS.

En nuestra experiencia como estudiantes, un aspecto positivo ha sido comprobar que modelos simples como la regresión logística —aunque menos sofisticados que COMPAS— resultan más fáciles de auditar y comprender. Esto nos hace pensar que:

1. Podría ser útil mantener **modelos de referencia básicos** como benchmarks de transparencia.

2. Sería interesante establecer **revisiones periódicas de sesgo y precisión**, quizá mediante dashboards accesibles para jueces y fiscales.

3. La **formación en alfabetización algorítmica** para los jueces, fiscales y defensores. Deberían recibir formación básica sobre el funcionamiento, las limitaciones y los sesgos potenciales de los algoritmos predictivos. Esta capacitación permitiría una interpretación crítica de las puntuaciones de riesgo, evitando una confianza ciega en la tecnología.

4. La decisión final debería combinar siempre **evaluación humana y modelo automático**, evitando que el algoritmo se convierta en el único criterio.
5. **Definir protocolos de transparencia y responsabilidad institucional.** Cada entidad que adopte un algoritmo debería establecer mecanismos claros de rendición de cuentas: quién valida los modelos, quién audita los resultados y cómo se comunican las limitaciones al público.

En ningún caso consideramos estas medidas como soluciones cerradas, sino como **ideas exploratorias** que creemos que pueden enriquecer la discusión.

En definitiva, nuestro análisis nos ha permitido constatar que los algoritmos predictivos como COMPAS no pueden evaluarse de forma aislada, sino en relación con la calidad y representatividad de los datos de partida. Creemos que la principal lección aprendida es sencilla pero fundamental: **sin datos justos y coherentes, ningún algoritmo podrá ser verdaderamente justo**. Nuestras propuestas no pretenden ofrecer soluciones cerradas, sino abrir el debate sobre cómo mejorar la transparencia y la equidad en la aplicación de estas herramientas dentro del sistema judicial.

11. BIBLIOGRAFÍA Y REFERENCIAS

- https://www.science.org/doi/10.1126/sciadv.aao5580
- https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf
- https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf
- https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- https://stanfordrewired.com/post/data-and-discretion
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Algorithmic fairness datasets: the story so far, Alessandro Fabris and Stefano Messina and Gianmaria Silvello and Gian Antonio Susto,2022.
- ProPublica's COMPAS Data Revisited, M. Barenstein, 2019
- https://cyber.harvard.edu/events/2018/luncheon/03/Dressel
- https://www.youtube.com/watch?v=G0OE8p-fc10&t=318s →(Conversatorio dado en la Universidad de Harvard sobre la precisión y equidad de algoritmos predictivos, usando el sistema COMPAS como ejemplo).
- https://www.youtube.com/watch?v=p-82YeUPQh0 (TED talk que alerta sobre los riesgos de los algoritmos predictivos como COMPAS, que pueden amplificar sesgos raciales y afectar injustamente el proceso judicial).

- https://www.youtube.com/watch?v=iR-xjr_6fGo&t=91s (Explicación sobre como reducir los sesgos en programas de ML).
- https://www.youtube.com/watch?v=f7oro8hgNZg&t=3s (Un ejemplo más de consejos para evitar los Sesgos).
- https://www.youtube.com/watch?v=2YI7_EdbEtY (Conversatorio de Joy Buolamwini. Fundadora de la Liga de la Justicia Algorítmica)
- https://www.ajl.org/ (Web oficial de la Liga de la Justicia Algorítmica- Fuente de inspiración)
- https://fairmlbook.org/ (Fairness and Machine Learning)

12. ANEXOS

12.1 DICCIONARIO:

- **COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)**: Algoritmo utilizado en EE.UU. para predecir riesgo de reincidencia.

- **Reincidencia**: En Estados Unidos, la reincidencia, que es la tendencia de un individuo a volver a cometer un delito después de haber sido condenado y sancionado, se aborda a través de diversas leyes y políticas. No hay un período de tiempo único para definir la reincidencia, ya que depende del delito, el historial delictivo del individuo y las leyes específicas del estado o federal.

- **Tasa base de criminalidad:** La *tasa base de criminalidad* hace referencia a la proporción real de delitos cometidos dentro de un grupo poblacional específico en relación con el total de individuos de ese grupo en la sociedad. En el contexto del proyecto, la ausencia de estas tasas base por etnia en el dataset de ProPublica impide contextualizar adecuadamente las diferencias observadas en las tasas de arresto o reincidencia. Esto limita la capacidad de distinguir si las disparidades entre grupos reflejan verdaderas conductas delictivas o responden a sesgos en la actuación policial y judicial.

- **Puntuación Compas**: Valor de 1 a 10 asignado por COMPAS.
    - 1–4 → Bajo riesgo
    - 5–7 → Riesgo medio
    - 8–10 → Alto riesgo

- **Sesgo Algorítmico**: Diferencias sistemáticas en la predicción de reincidencia según grupo demográfico (ej. etnia o género).

- **Felony** (F): Son los delitos graves. Se castigan con penas de prisión de más de un año, y en algunos casos incluso con cadena perpetua o pena de muerte. Ejemplos: homicidio, violación, robo a mano armada, tráfico de drogas a gran escala. Suelen juzgarse en tribunales estatales o federales de mayor nivel.

- **Misdemeano**r (M): Son los delitos menores. Se castigan normalmente con multas, libertad condicional o penas de cárcel inferiores a un año, muchas veces en cárceles locales. Ejemplos: hurtos menores, conducir bajo los efectos del alcohol, alteración del orden público.

- **Métricas de Evaluación**: **TPR (True Positive Rate / Sensibilidad)**: % de reincidentes correctamente clasificados. **FPR (False Positive Rate)**: % de no reincidentes clasificados erróneamente como reincidentes. **ROC-AUC**: Medida de discriminación del modelo.

12.1.2 CONCEPTOS DE MACHINE LEARNING

- **Regresión Logística**: Modelo de clasificación usado como alternativa a COMPAS para predecir reincidencia.

- **Modelo Simple (baseline)**: Modelo implementado por el grupo para comparar frente a COMPAS, evitando introducir sesgos demográficos.
  KPIs de evaluación: Precisión, Recall, FPR, TPR, etc., desglosados por grupos demográficos.

12.2 CÓDIGO RELEVANTE (NOTEBOOKS, FUNCIONES)

12.3 TABLAS O GRÁFICOS COMPLEMENTARIOS

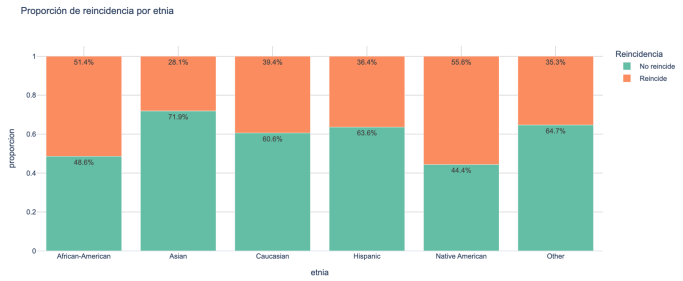


Fig. 22. Proporción de reincidencia por etnia.

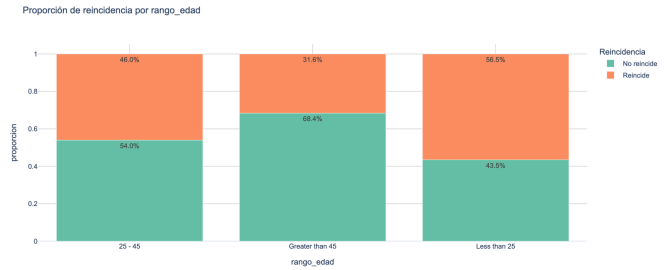


Fig. 23. Proporción de reincidencia por rango de edad.

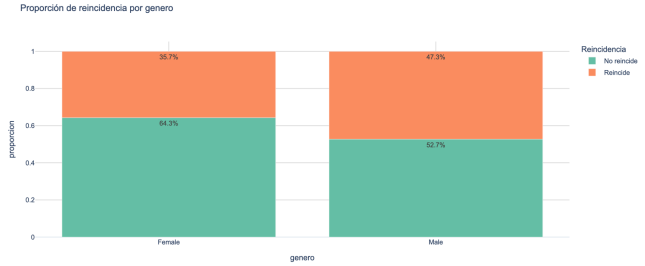


Fig. 24. Proporción de reincidencia por género.

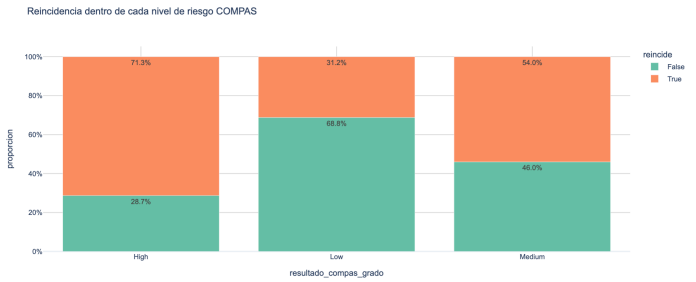


Fig. 25. Reincidencia dentro de cada nivel de riesgo COMPAS.

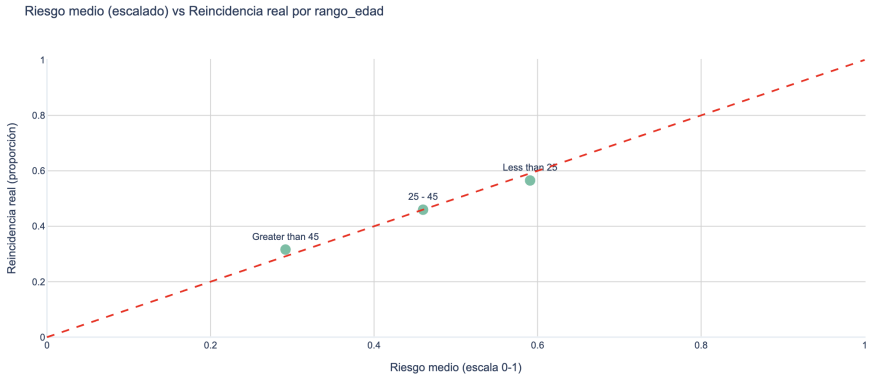


Fig. 26. Riesgo medio escalado vs Reincidencia real por rango de edad