



BOURBAKI  
COLEGIO DE MATEMÁTICAS

## Causalidad en Machine Learning IV

El cálculo Do y la solución a la paradoja de Simpson: el criterio Back-Door



# Índice

01. Introducción \_\_\_\_\_ pág. 02

02. La paradoja de Simpson \_\_\_\_\_ pág. 03

01. ¿Cómo observar la causalidad? — pág. 04

02. Un ejemplo con bicicletas — pág. 05

03. La paradoja de Simpson — pág. 07

04. De Sewall Wright a Judea Pearl — pág. 08

03. El cálculo do y la paradoja de Simpson

pág. 09

01. Prestamos vs línea de crédito — pág. 09

02. Teorema de imposibilidad — pág. 12

04. Lectura de referencia \_\_\_\_\_ pág. 18

05. Caso de uso: Equidad algorítmica — pág. 20

# 01 Introducción

La causalidad es un análogo asimétrico del concepto de correlación, esta última es una de las relaciones más importantes en Machine Learning. En este curso, buscamos familiarizar al Científico de Datos con los fundamentos y las ideas detrás del estudio matemático de la causalidad. Por medio de casos de uso reales practicaremos el uso de modelos causales para el desarrollo de modelos más confiables.

El contenido del curso se divide en seis módulos:

- I. Interpretabilidad, Inferencia Bayesiana y el operador Do
- II. Ensayos aleatorizados: tests A/B, Simulación y RCT
- III. Modelos Causales, D separación y los Axiomas de la Causalidad
- IV. El cálculo Do y la solución a la paradoja de Simpson: el criterio Back Door
- V. Propensity Score: ventajas y controversias en el Do-Calculus
- VI. Ortogonalización y Doble Machine Learning

 El repositorio de esta semana está disponible en [este Link](#).

## 02 La paradoja de Simpson



A medida que un científico de datos gana experiencia implementando modelos matemáticos, descubre la enorme necesidad de no solo comprender la correlaciones entre variables, sino también un análisis causal entre las variables explicativas y las variables objetivo.

Desafortunadamente, la mayoría de las técnicas estadísticas clásicas, e incluso los modelos de machine learning, no están diseñados para identificar relaciones causales entre las características y las observaciones históricas de las variables que deseamos predecir.

El análisis de la causalidad es un amargo crítico del enfoque actual para la Inteligencia Artificial mediante redes neuronales costosamente entrenadas en bases de datos gigantescas pues, bien entendido, un modelo causal difícilmente requiere grandes volúmenes de datos ni conexiones complejas entre ellos como las redes neuronales profundas.

No obstante, nos gustaría insistir en que no hay alguna razón conocida por nosotros por la que el enfoque de causalidad sea incompatible con el de las redes neuronales profundas.

## ¿Cómo observar la causalidad?

---

En la actualidad es indiscutible que el fumar es una causa del cáncer de pulmón; sin embargo en 1958 el célebre genetista y estadístico Ronald A. Fisher publicó un artículo, ni más ni menos que en la revista *Nature* dudando sobre esta relación causal y sugiriendo que podría ser únicamente una correlación entre los datos.

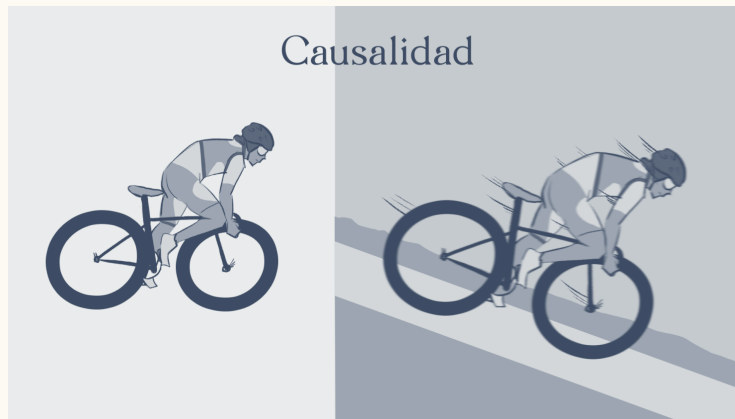
El artículo se titula *Cancer and Smoking* y es uno de nuestros ejemplos favoritos para explicar lo difícil que puede ser establecer una relación causal. Imaginémonos intentando contradecir a un afamado científico de Cambridge University que además publicó sus estudios en la revista científica más prestigiosa.

Entonces, ¿Cómo justificar, matemáticamente la causalidad entre dos conjuntos de variables observadas  $X$  e  $Y$ ?

Como bien lo sabe el 100% de los científicos de datos o analistas de negocio, una alta correlación de Pearson no necesariamente implica causalidad. Incluso cuando las variables están en proporción directa, no significa que una de estas variables haya ocasionado a la otra.

## Un ejemplo con bicicletas

---



Consideremos tres variables aleatorias  $X$ ,  $Z$  e  $Y$  que representen la siguiente información sobre un conjunto de ciclistas que están descendiendo de una montaña:

- La variable  $X$  registra el esfuerzo del ciclista al pedalear. Por comodidad, denotaremos por  $-X$  a un pedaleo muy ligero.
- La variable  $Y$  representa una alta velocidad alcanzada por el ciclista.
- La variable  $Z$  denota una gran pendiente negativa de la montaña que los ciclistas descienden, mientras que  $-Z$  también es una pendiente negativa pero no tan pronunciada. Es decir,  $Z < -Z < 0$ .

Es bastante intuitivo imaginar que una base de datos con la información anterior satisface las siguientes observaciones estadísticas.

- Las variables  $X$  e  $Y$  están positivamente correlacionadas. Digamos que su coeficiente al entrenar una regresión es positivo:  $A > 0$ . Esto signi-

fica que entre más se esfuerza por pedalear el ciclista  $X$ , mayor será la probabilidad de alcanzar una velocidad alta.

$$\mathbb{P}(Y | X) > \mathbb{P}(Y | -X)$$

- Las variables  $Z$  e  $Y$  están positivamente correlacionadas. Su coeficiente al entrenar una regresión es positivo:  $B > 0$ . De nuevo, entre mayor sea la pendiente, más alta será la probabilidad de que el ciclista alcance una gran velocidad.

$$\mathbb{P}(Y | Z) > \mathbb{P}(Y | -Z)$$

- Las variables  $Z$  y  $X$  están negativamente correlacionadas. Su coeficiente al entrenar una regresión es igual a  $C < 0$ .

Supongamos que deseamos investigar la relación causal entre el esfuerzo al pedalear y la velocidad a la que está rodando el ciclista. Parecería natural suponer que la causa de la velocidad es su esfuerzo al pedalear. Desafortunadamente aunque nos gustaría llegar rápidamente a esta conclusión, los datos podrían no ayudarnos.

Un ejemplo perfecto de esto es la llamada paradoja de Simpson, que es una de las observaciones más agudas sobre relaciones causales que existen en estadística, y sólo fue completamente resuelta por el padre del análisis matemático de la causalidad, Judea Pearl.

## La paradoja de Simpson

---

La paradoja de Simpson postula lo siguiente: A pesar de que las variables  $X$  e  $Y$  de nuestro ejemplo están correlacionadas positivamente, el signo de esta correlación podría revertirse al segmentar los datos de acuerdo a la otra variable explicativa. Para los ciclistas que recorren una pendiente muy pronunciada  $Z$ , una correlación negativa entre  $X$  y  $Y$  significa que mientras menos pedaleen, mayor será la probabilidad de lograr una alta velocidad debido al impulso natural del terreno.

$$\mathbb{P}(Y | X, Z) < \mathbb{P}(Y | -X, Z)$$

¿Es esto posible?, ¿Podría ocurrir un cambio en el signo de la correlación cuando consideramos a los ciclistas que descienden en pendientes poco pronunciadas? Es decir, ¿se cumple la siguiente desigualdad?

$$\mathbb{P}(Y | -X, Z) < \mathbb{P}(Y | -X, -Z)$$

Existen numerosos casos reales en los que ambas preguntas se contestan positivamente lo cual es sumamente contra-intuitivo y no fue sino hasta el surgimiento del famoso cálculo Do de Judea Pearl que se comprendió cabalmente esta paradoja.



## De Sewall Wright a Judea Pearl

---

En el año de 1921 el genetista Sewall Wright publicó un artículo titulado *Correlation and causation* el cual sería la antesala de su artículo en 1934 titulado *The method of Path coefficients*. En estos trabajos Wright propone un método para analizar matemáticamente la existencia de relaciones causales entre variables observadas.

Aunque la paradoja de Simpson no puede resolverse por completo utilizando el método de Wright, muestra que para el caso lineal, es posible utilizar la correlación parcial como herramienta para acercarnos a una solución la paradoja de Simpson.

Mediante el análisis de Wright se puede concluir que condicionando con la pendiente del descenso, la correlación parcial entre el esfuerzo al pedalear y la velocidad del ciclista se puede calcular de la siguiente manera:

$$\text{Corr}(X, Y|Z) = A + BC$$

Si el efecto negativo de  $Z$  sobre  $X$  es menor que  $-(A/B)$  entonces es mejor suponer que existe una correlación negativa entre  $X$  e  $Y$  aunque un análisis entre ellas no lo sugiera.

## 03 El cálculo de y la paradoja de Simpson

El siguiente planteamiento supone la necesidad de un banco de evaluar la preferencia de dos de sus productos. Los dos productos contendientes son:

$X = 1$ : Prestamos personales

$X = -1$  Aumentos en la línea de crédito

Esta oferta será lanzada unicamente entre los clientes que ya tienen un crédito activo con el banco:

$Z = 0$  Clientes que tienen una hipoteca

$Z = 2$  clientes que tienen un crédito automotriz.

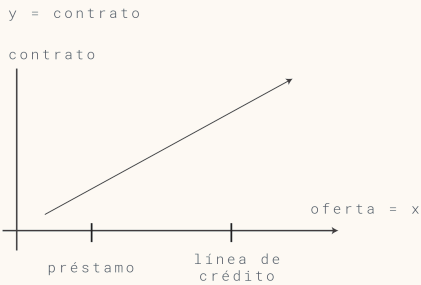
### Prestamos vs línea de crédito

---

En un primer estudio, los clientes en general tuvieron las siguientes preferencias. Del total de la población, 78% declaró que prefería la oferta de un préstamo personal y 86% declaró que prefería la oferta de aumentar su línea de crédito.

Si nuestro trabajo es que los clientes acepten firmar un contrato con el banco ( $Y$ ), la gráfica de la derecha ilustra la probabilidad de éxito de que los dos

OFERTA	CONTRATO
<div>Préstamo</div> <div></div>	<div>78%</div> <div>No</div>
<div>Línea de Crédito</div> <div></div>	<div>86%</div> <div>No</div>



tipos de clientes acepten la oferta:

$$\mathbb{P}(Y|X = 1) < \mathbb{P}(Y|X = -1)$$

por lo que según este primer estudio, sería natural pensar que la respuesta correcta es ofrecer a todos un aumento en la línea de crédito.

A continuación separamos los datos respecto a los dos tipos de clientes.

	CLIENTE	OFERTA	CONTRATO
A <sub>1</sub>	Hipoteca 	Préstamo 	78%
a <sub>1</sub>	Auto 		
A <sub>2</sub>	Hipoteca 	Línea de Crédito 	No
a <sub>2</sub>	Auto 		
B <sub>1</sub>	Hipoteca 	Línea de Crédito 	86%
b <sub>1</sub>	Auto 		
B <sub>2</sub>	Hipoteca 	Línea de Crédito 	No
b <sub>2</sub>	Auto 		

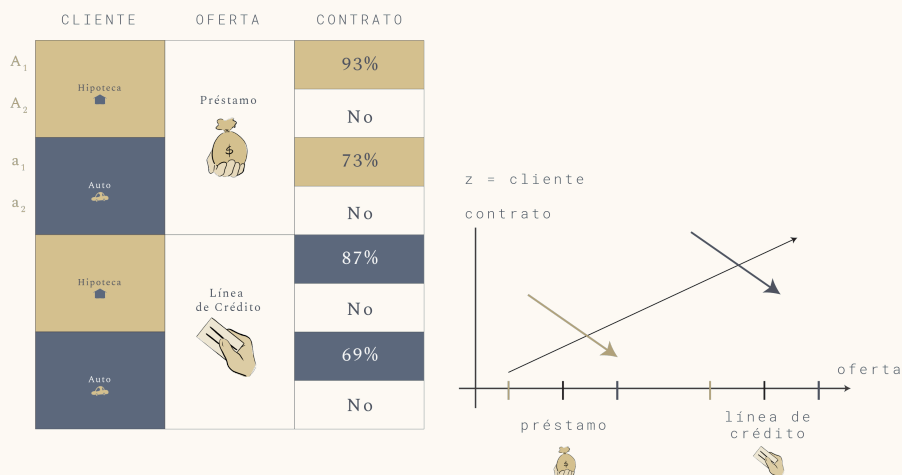
Por ejemplo:

A <sub>1</sub> = 81	B <sub>1</sub> = 234
a <sub>1</sub> = 87	b <sub>1</sub> = 240
A <sub>2</sub> = 192	B <sub>2</sub> = 55
a <sub>2</sub> = 263	b <sub>2</sub> = 80

Con los datos numéricos verificamos los porcentajes iniciales.

$$78\% = \frac{A_1 + a_1}{A_2 + a_2} \quad \frac{B_1 + b_1}{B_2 + b_2} = 86\%$$

Consideremos el siguiente reordenamiento de los datos: Juntemos a todas las personas que tienen hipotecas y observemos que el  $93\% = A_1 / A_2$  de ellas prefiere la oferta del préstamo. Mientras que el  $73\% = a_1 / a_2$  las personas con crédito automotriz también prefieren la oferta del préstamo. Por otro lado, el  $87\% = B_1 / B_2$  de las personas con hipoteca prefieren la oferta de aumentar línea de crédito, así como el  $69\%$  de las personas con credito automotriz.



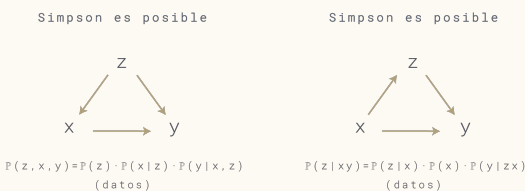
Observemos que los nuevos porcentajes han invertido la desigualdad 01. De hecho, ahora la probabilidad de firmar un nuevo contrato se ha segmentado como en la gráfica de la derecha.

$$\begin{aligned} \mathbb{P}(Y|X=1, Z=0) &> \mathbb{P}(Y|X=1, Z=2); & 93\% = \frac{A_1}{A_2} &> \frac{B_1}{B_2} = 87\% \\ \mathbb{P}(Y|X=-1, Z=0) &> \mathbb{P}(Y|X=-1, Z=2); & 73\% = \frac{a_1}{a_2} &> \frac{b_1}{b_2} = 69\% \end{aligned}$$

La pregunta natural es ¿cómo proceder ahora? Por un lado los datos nos dicen que el mejor producto a ofrecer es el aumento a la línea de crédito. Pero segmentando a la población tenemos que ambos grupos prefieren que se les ofrezca el préstamo. En las siguientes secciones abordaremos la solución a este problema.

## Teorema de imposibilidad

Como ya lo hemos visto en la sección anterior, en algunos casos estaremos en la encrucijada entre elegir si condicionamos con las variables  $Z$  o no para calcular una relación entre un tratamiento  $X$  y la cura  $Y$ . En este capítulo que está dividido en dos partes explicaremos cómo los modelos causales y el operador do que hemos aprendido en el módulo pasado nos permiten resolver satisfactoriamente la paradoja de Simpson.



## 02.1 Solución I: un mundo sin la paradoja de Simpson

---

Uno de los grandes logros del trabajo de Judea Pearl es haber encontrado un modelo matemático en el que la paradoja de Simpson no es cierta como veremos a continuación en el siguiente teorema.

**Theorem 02.1.** *Supongamos que un conjunto de variables  $Z, X, Y$  donde  $X$  es binaria satisfacen lo siguiente, la notación  $-X$  significa que la variable  $X$  toma el otro de los valores binarios:*

$$\mathbb{P}(Z|do(X)) = \mathbb{P}(Z) = \mathbb{P}(Z|do(-X))$$

*Entonces la paradoja de Simpson es imposible cuando las probabilidades condicionales están calculadas con el operador do.*

## 02.2 Solución II: El criterio de Back-door

---

En esta sección abordaremos el desafío central para resolver la paradoja de Simpson desde un punto de vista práctico: demostrar que es posible tomar la decisión correcta cuando ocurre una inversión en las asociaciones estadísticas.

El utilizará el cálculo do el cuál es una herramienta algebraica que permite determinar el efecto causal de una variable sobre otra, identificando cuáles variables deben ser utilizadas para condicionar con el objetivo de estimar co-

rectamente el efecto causal.

El criterio de back-door nos ayudará a distinguir si debemos de condicionar o no mediante un conjunto de variables  $Z$  para encontrar entre caminos causales (deseados) y espurios (indeseados) entre dos variables  $X$  e  $Y$ .

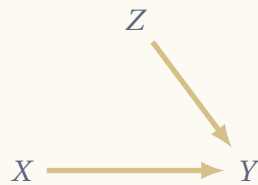
**Definition 02.1.** Sean  $X$  e  $Y$  dos variables dentro de un modelo causal y sea  $Z$  un conjunto de variables dentro del mismo modelo. Diremos que  $Z$  satisface el **criterio back-door** para la causalidad entre  $X$  e  $Y$  cuando se cumpla lo siguiente:

1. Ninguna variable en  $Z$  es descendiente de  $X$ .
2. Para cualquier camino  $\tau = (X, v_2, \dots, Y)$  entre  $X$  e  $Y$  tal que la flecha  $(v_2, X)$  es el inicio del camino, se cumple que  $Z$  lo está bloqueando.

Este criterio es equivalente a la hipótesis de independencia que hicimos entre las variables contrafactuales y el tratamiento  $X$  condicionado con las variables  $Z$ .

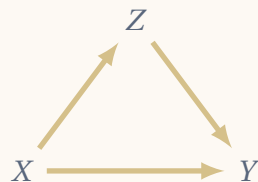
Es relativamente sencillo calcular si este criterio se cumple o no para una red bayesiana. Practiquemos este criterio con los siguientes ejemplos:

**Example 02.2.** *En este primer caso sí se cumple el criterio pues el único camino entre  $X$  e  $Y$  trivialmente satisface la segunda condición en la definición anterior.*

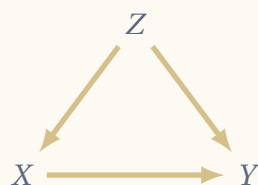


*Este ejemplo es muy relevante pues estamos observando independencia entre las variables  $X$  y  $Z$  por lo cual estamos bajo las hipótesis del teorema Sure-Thing de la sección anterior lo cual significará que la paradoja de Simpson no es posible. Observemos por último que la independencia entre  $X$  y  $Z$  es un indicio del caso ideal de un experimento aleatorizado.*

**Example 02.3.** *En este segundo caso no se cumple el criterio de Back-door debido a la primera condición.*



**Example 02.4.** *en este tercer caso sí se cumple el criterio de Back-door debido a la segunda condición.*



La importancia del criterio back-door está justificada por el siguiente teorema:



**Theorem 02.5.** *Supongamos que un conjunto con  $K$  variables  $Z$  satisfacen el criterio de back-door para la causalidad entre otras dos variables  $X$  e  $Y$ . Entonces la relación causal entre  $X$  e  $Y$  se puede calcular de la siguiente manera:*

$$\mathbb{P}(Y|do(X)) = \mathbb{P}(Y|X, Z_1)\mathbb{P}(Z_1) + \dots + \mathbb{P}(Y|X, Z_K)\mathbb{P}(Z_K)$$

La conclusión de este criterio la hemos estado anhelando desde el módulo II con las variables contrafactuales pues significa que podemos medir causalidad utilizando únicamente datos observados.

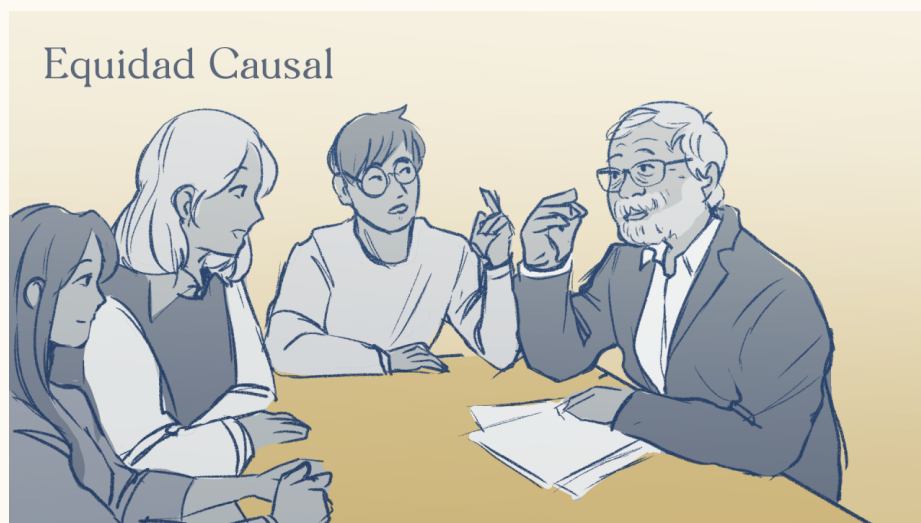
Regresando a los tres ejemplos anteriores podemos concluir lo siguiente:

1. En el primer ejemplo ya mencionamos que la paradoja de Simpson ni siquiera será observada estadísticamente, esto significa que el efecto causal entre  $X$  e  $Y$  será lo mismo con o sin condicionar con la variable  $Z$ .
2. El segundo ejemplo como ya lo mencionamos no cumple el criterio de Back-door por lo que para calcular la relación causal entre  $X$  e  $Y$  no debemos de utilizar la fórmula en la conclusión del teorema anterior, es decir que no debemos de condicionar con la variable  $Z$ .
3. En el tercer ejemplo el conjunto  $Z$  sí cumple con el criterio Back-Door por lo que para calcular el efecto causal entre  $X$  e  $Y$  sí debemos de condicionar con la variable  $Z$ .

Como lo vemos, gracias a este criterio, es posible determinar en qué casos

conviene condicionar en  $Z$  y cuándo no hacerlo. El resultado es una guía confiable para decidir si los datos desagregados o los agregados proporcionan la respuesta correcta.

## 04 Lectura de referencia



Todos los que trabajamos en algún aspecto relacionado con Machine Learning sabemos lo delicado que es suponer que los modelos matemáticos van a comportarse por sí solos bajo los criterios morales, legales o éticos que nos gustaría observar en las personas a nuestro alrededor. Esto representa un problema bastante complicado cuando el uso de la Inteligencia Artificial se ha convertido en un aspecto tan importante para nuestra vida diaria, como en sistemas automatizados de toma de decisiones en áreas como salud, justicia, educación y finanzas. Si bien estos sistemas pueden parecer neutrales, frecuentemente reproducen o incluso amplifican las desigualdades sociales presentes en los datos de entrenamiento.

El artículo de referencia plantea una forma de analizar la justicia en los algoritmos que no solo refleje los datos, sino que intente entender las causas

de las desigualdades. A esto le llaman el Problema Fundamental del Análisis de Justicia Causal (FPCFA), y consiste básicamente en comprender de dónde vienen las diferencias en los resultados. Se trata de distinguir si son producto de condiciones justas o de mecanismos que generan discriminación. Para ello, se apoyan en conceptos legales como el trato desigual y el impacto desigual, y su meta es medir qué parte de esa desigualdad es verdaderamente injusta. Con este enfoque, buscan superar las limitaciones de los métodos estadísticos tradicionales, que no pueden decirnos si una diferencia en los resultados es causada directamente por una injusticia. Además, ofrecen herramientas prácticas como el Mapa de Justicia y el Fairness Cookbook, para ayudar a quienes desarrollan modelos a evaluar y corregir posibles sesgos en sus sistemas.

 El artículo de referencia, Causal Fairness Analysis se puede encontrar en [este link](#).

## 05 Caso de uso: Equidad algorítmica




ProPublica, organización de periodismo de investigación con sede en New York, analizó el uso del algoritmo COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), empleado por varios tribunales de EE.UU. para predecir la probabilidad de reincidencia criminal de personas previamente acusadas de cometer algún crimen.

Los investigadores se dieron a la tarea de analizar los scores asignados a más de 7000 personas de Broward County, Florida arrestados entre 2013 y 2014. Compararon las predicciones del algoritmo con su comportamiento real durante los dos años posteriores a su arresto.

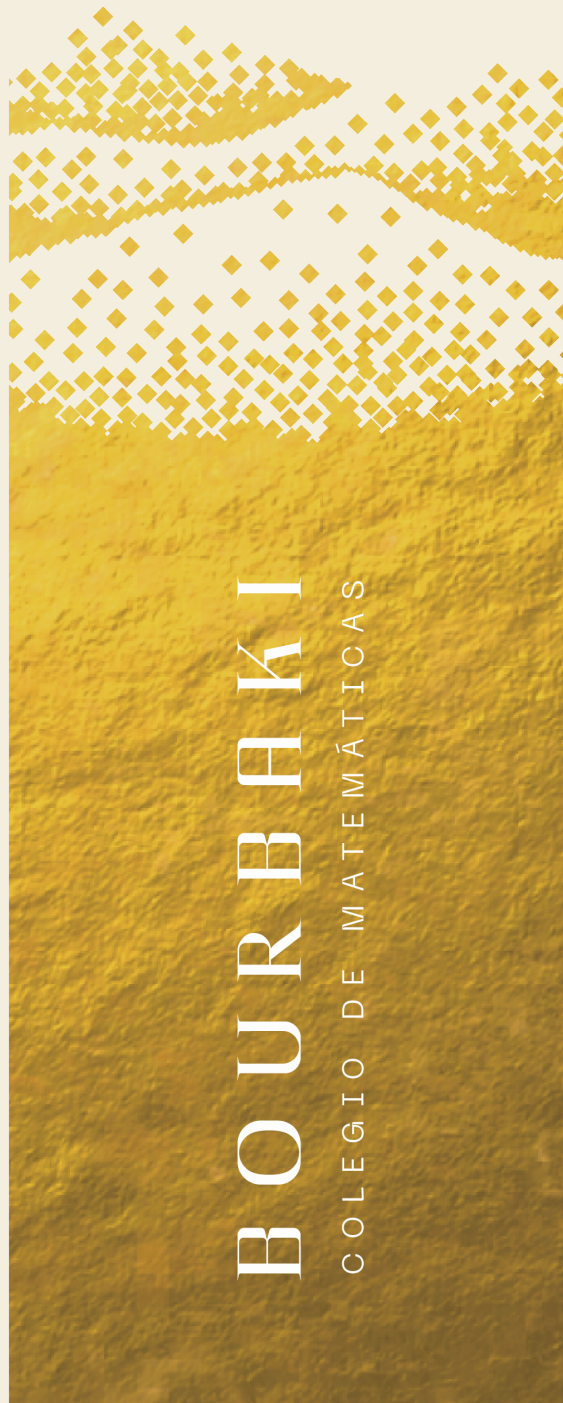
Encontraron que el algoritmo acertaba en aproximadamente 61 % de los casos y que las tasas de error eran similares entre blancos y negros, sin embargo,

notaron que en los falsos positivos (etiquetados como personas de alto riesgo que en realidad no reincidieron), había casi el doble de personas afroamericanas (44.9%) que de personas blancas (23.5%); mientras que en los falsos negativos (personas etiquetadas como seguras que sí reincidieron), las personas blancas eran etiquetadas más frecuentemente.

Este es solo un ejemplo de la importancia del análisis causal para comprender disparidades y distinguir correlaciones espurias de efectos puramente discriminatorios, pero sobre todo, para diseñar modelos más justos que no impacten sobre grupos vulnerables.

 El artículo Machine Bias se puede encontrar en [este link](#).

colegio-bourbaki.com  
+52 56 2141 7850  
info@colegio-bourbaki.com



BOURBAKI  
COLEGIO DE MATEMÁTICAS