presentado en

Pista 2

Investigación sobre los sesgos

Lo que veremos en este módulo...

1. ¿Qué es el sesgo en la IA?
2. **Caso práctico:** ¿Puede ChatGPT superar una prueba de discriminación laboral?
3. Investigación del sesgo de la IA a lo largo de su ciclo de vida.
4. **Ejercicio interactivo:** Investigar el sesgo de un modelo real de IA.

Lo que veremos en este módulo...

1. **¿Qué es el sesgo en la IA?**
2. **Caso práctico:** ¿Puede ChatGPT superar una prueba de discriminación laboral?
3. Investigación del sesgo de la IA a lo largo de su ciclo de vida.
4. **Ejercicio interactivo:** Investigar el sesgo de un modelo real de IA.

Las empresas y el gobierno prometen que la IA será objetiva y justa

Becoming truly data driven is an ambition of the city of Rotterdam.

In order to more accurately identify illegitimate welfare recipients and increase compliance by both the citizen and the city overall, they took a new, sophisticated data-driven approach.

ADVANCED ANALYTICS + **MACHINE LEARNING** = **UNBIASED CITIZEN OUTCOMES**

Un informe técnico de la multinacional Accenture promociona su trabajo en un modelo de aprendizaje automático para puntuar a los beneficiarios de ayudas sociales en la ciudad de Róterdam.

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

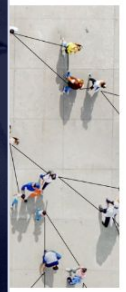
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Prediction: Bias

Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them



SURVEILLANCE

Suspicion Machines

Unprecedented experiment on welfare surveillance algorithm reveals discrimination

¿Por qué?

Fundamentalmente:
Si entrenamos un sistema de IA con datos **sesgados**, sus predicciones o su comportamiento **reproducirá** esos sesgos.

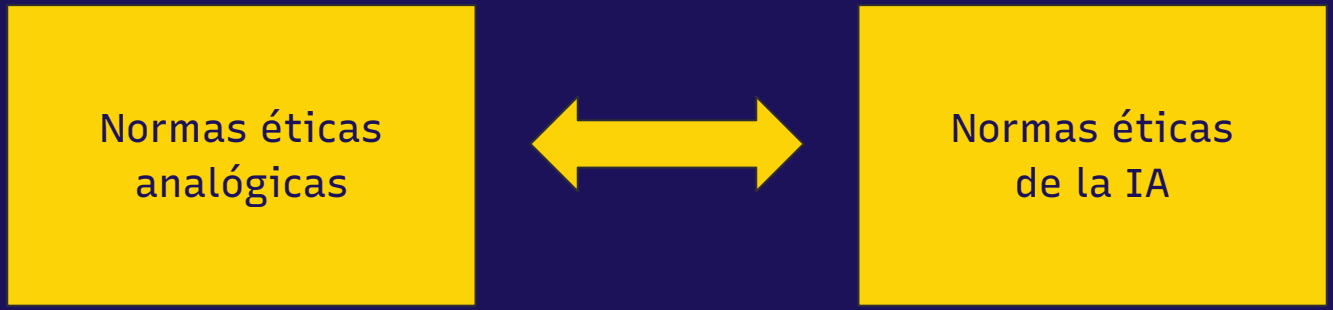
Pero, ¿qué queremos decir realmente cuando decimos que un sistema de IA es «sesgado»?

¿Es sesgado un modelo que utiliza la edad para predecir el riesgo de cáncer?

Lo que veremos en este módulo...

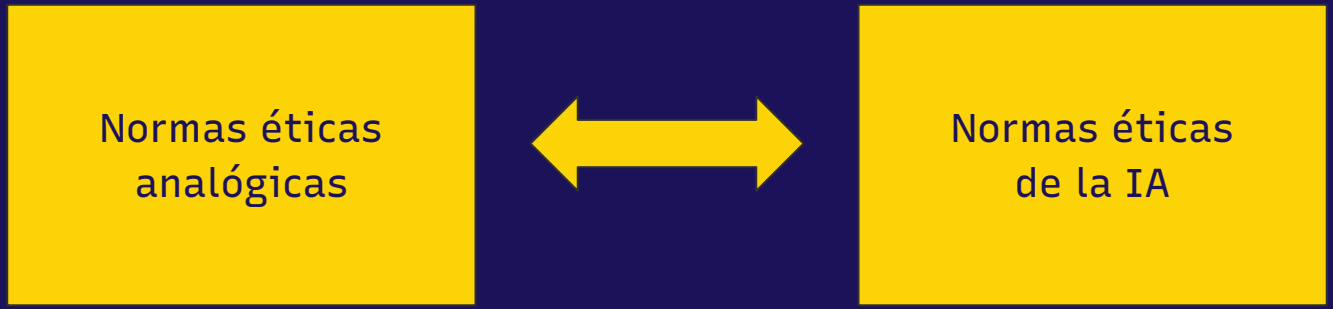
1. ¿Qué es el sesgo en la IA?
2. **Caso práctico: ¿Puede ChatGPT superar una prueba de discriminación laboral?**
3. Investigación del sesgo de la IA a lo largo de su ciclo de vida.
4. **Ejercicio interactivo:** Investigar el sesgo de un modelo real de IA.

La IA no existe de forma aislada



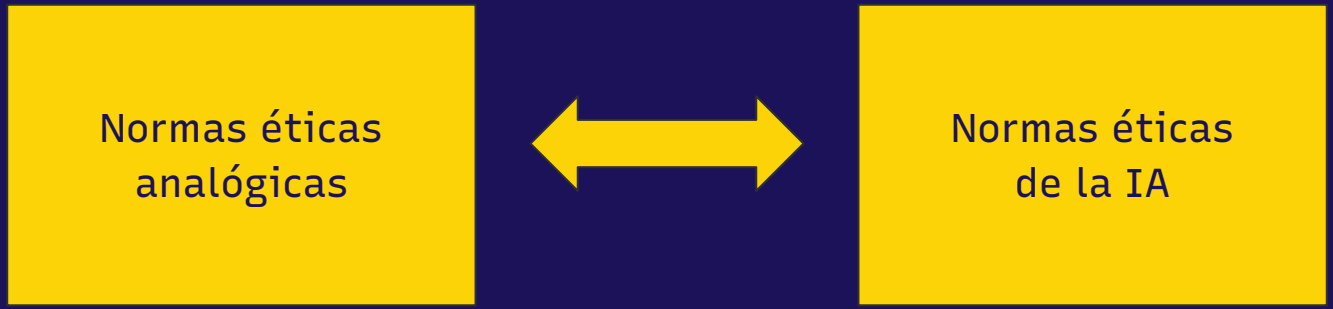
- Los avances en el aprendizaje automático no suponen una desviación fundamental de los sistemas analógicos (personas que toman decisiones) → las consideraciones de equidad de la IA se aplican a todas las tecnologías.
- Las normas éticas analógicas pueden utilizarse para interrogar a los sistemas de aprendizaje automático **y viceversa**.

¿Qué queremos decir con esto?



- Discriminación laboral: personas con la misma cualificación para un puesto de trabajo no son contratadas por motivos de raza, género, edad u otras características protegidas.

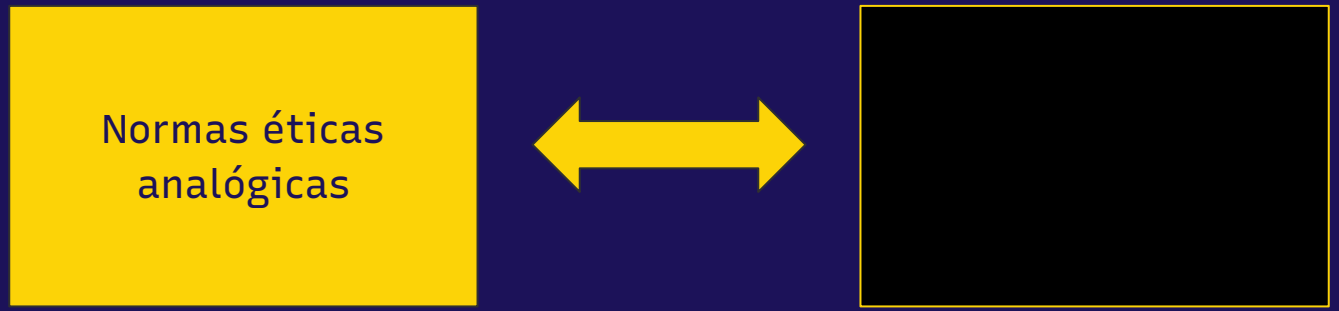
¿Qué queremos decir con esto?



- ¿Cómo estudian los académicos la discriminación laboral en el mundo analógico?
- Envían EXACTAMENTE EL MISMO CV/currículum y solo cambian el nombre. Ven quién recibe una llamada.
- Resultado: los currículums con nombres racialmente distintos reciben menos llamadas.

¿Supera ChatGPT las pruebas analógicas de discriminación laboral?

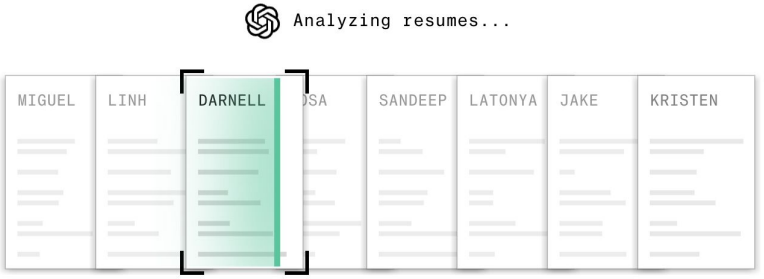
¿Qué pasaría si probáramos ChatGPT para detectar la discriminación laboral?



- La misma prueba analógica, pero ahora con ChatGPT:
- Pídele a ChatGPT que clasifique los MISMOS CV/currículums, pero con nombres diferentes y racialmente distintos.

Resultado

Bloomberg

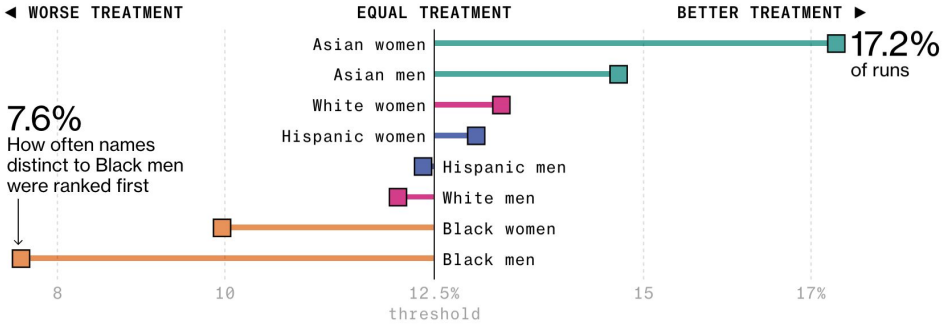


OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

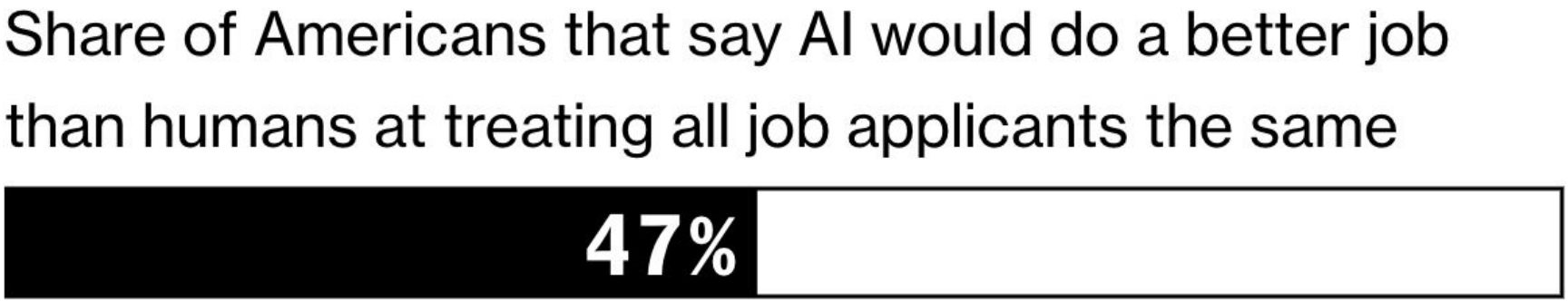
By Leon Yin, Davey Alba and Leonardo Nicoletti
for **Bloomberg Technology** + **Equality**

7 March 2024



Source: Bloomberg Analysis of OpenAI's GPT-3.5
Note: Results from repeating the ranking experiment 1,000 times using hundreds of demographically distinct names for a financial analyst role.

Por qué es importante



Lo que veremos en este módulo...

1. ¿Qué es el sesgo en la IA?
2. **Caso práctico:** ¿Puede ChatGPT superar una prueba de discriminación laboral?
3. **Investigación del sesgo de la IA a lo largo de su ciclo de vida.**
4. **Ejercicio interactivo:** Investigar el sesgo de un modelo real de IA.

¿En qué puntos del «ciclo de vida de la IA» pueden surgir sesgos?

¿Cómo podemos analizar el sesgo a lo largo del ciclo de vida de la IA?

¿El sistema de IA utiliza variables que son injustas, como la raza o el género de una persona? ¿Utiliza variables sustitutivas para estas características, como el código postal?

¿Contienen los datos de entrenamiento sesgos históricos? ¿Son representativos del mundo real?

¿La técnica de aprendizaje automático introduce aleatoriedad?

¿El sistema funciona igual de bien en todos los grupos?

¿Existe un impacto desigual sobre los grupos vulnerables?

¿La aplicación de un sistema de IA es en sí misma sesgada?

Una nota rápida sobre las variables proxy...

Variables (entradas) que se aproximan a otra variable o característica.

Variables de entrada

¿Utiliza el sistema de IA variables que son injustas, como la raza o el género de una persona? ¿Utiliza variables proxy para estas características, como el código postal?

Proxies comunes para la **raza** o el **origen étnico**:

- Código postal
- Apellido
- Estatus socioeconómico
- Composición del hogar

¿Qué tipo de materiales nos permiten comprobar si hay sesgos?

- **Hágalo usted mismo:** obtenga acceso a un modelo (por ejemplo, una clave OpenAI) y pruébelo usted mismo.
- **Informes internos:** Los gobiernos y las empresas auditan cada vez más sus propios modelos en busca de sesgos; intenta obtener los resultados.
- **De abajo hacia arriba:** trabaje con las comunidades para recopilar datos y/o testimonios de forma colaborativa.
- **Académicos:** Trabaja con académicos que estén desarrollando metodologías únicas o que puedan tener acceso a datos exclusivos.

Diferentes formas de pensar sobre el sesgo de la IA

Ciertos grupos no deben estar sobrerrepresentados o infrarrepresentados en las predicciones de un sistema de IA.

Ejemplo: un sistema de IA que selecciona de manera desproporcionada a mujeres para investigaciones por fraude es sesgado.

Un sistema de IA debe funcionar igual de bien en todos los grupos

Ejemplo: un sistema de reconocimiento facial debería funcionar igual de bien con tonos de piel más oscuros y más claros.

Los errores que comete un sistema de IA deben distribuirse de manera equitativa entre los grupos.

Ejemplo: un sistema de IA que predice erróneamente que los acusados negros serán futuros delincuentes en comparación con los acusados blancos es sesgado.

Tenga en cuenta que los debates sobre el sesgo de la IA no siempre son sencillos.

The Washington Post

This article was published more than **7 years ago**

MONKEY CAGE

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016 at 5:00 a.m. EDT

El sesgo de la IA tiene que ver, en última instancia, con el contexto, que esperamos que haya definido en su informe.

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In

Lo que veremos en este módulo...

1. ¿Qué es el sesgo en la IA?
2. **Caso práctico:** ¿Puede ChatGPT superar una prueba de discriminación laboral?
3. Investigación del sesgo de la IA a lo largo de su ciclo de vida.
4. **Ejercicio interactivo: Investigar el sesgo de un modelo real de IA.**

Investiguemos algunos sesgos algorítmicos del mundo real

- Un sistema de IA en la ciudad de Róterdam intenta predecir qué beneficiarios de ayudas sociales están cometiendo fraude.
- El sistema asigna a cada persona una puntuación de riesgo entre 0 y 1, siendo 1 el riesgo más alto de fraude.
- El sistema utiliza 315 variables de entrada para cada persona, incluyendo el idioma que hablan, su género y edad.
- El sistema se ha entrenado con datos de investigaciones anteriores que la ciudad ha llevado a cabo sobre los beneficiarios de ayudas sociales.

¿Hay alguna señal de alarma aquí?

Investiguemos algunos sesgos algorítmicos del mundo real

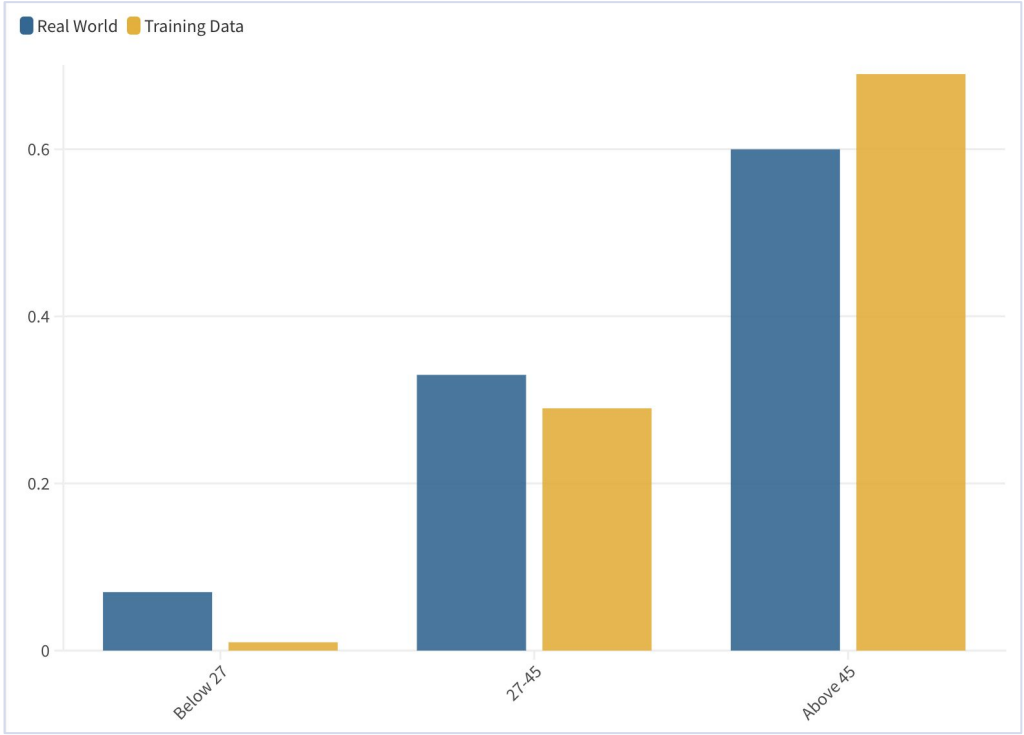
- Un sistema de IA en la ciudad de Róterdam intenta predecir qué beneficiarios de ayudas sociales están cometiendo fraude.
- El sistema asigna a cada persona una puntuación de riesgo entre 0 y 1, siendo 1 el riesgo más alto de fraude.
- El sistema utiliza 315 variables de entrada para cada persona, incluyendo el idioma que hablan, su género y edad.
- **El sistema se ha entrenado con datos de investigaciones anteriores que la ciudad ha llevado a cabo sobre los beneficiarios de ayudas sociales.**

¿Qué hay de los datos de entrenamiento?

Investiguemos los datos de entrenamiento

¿Notas algo?

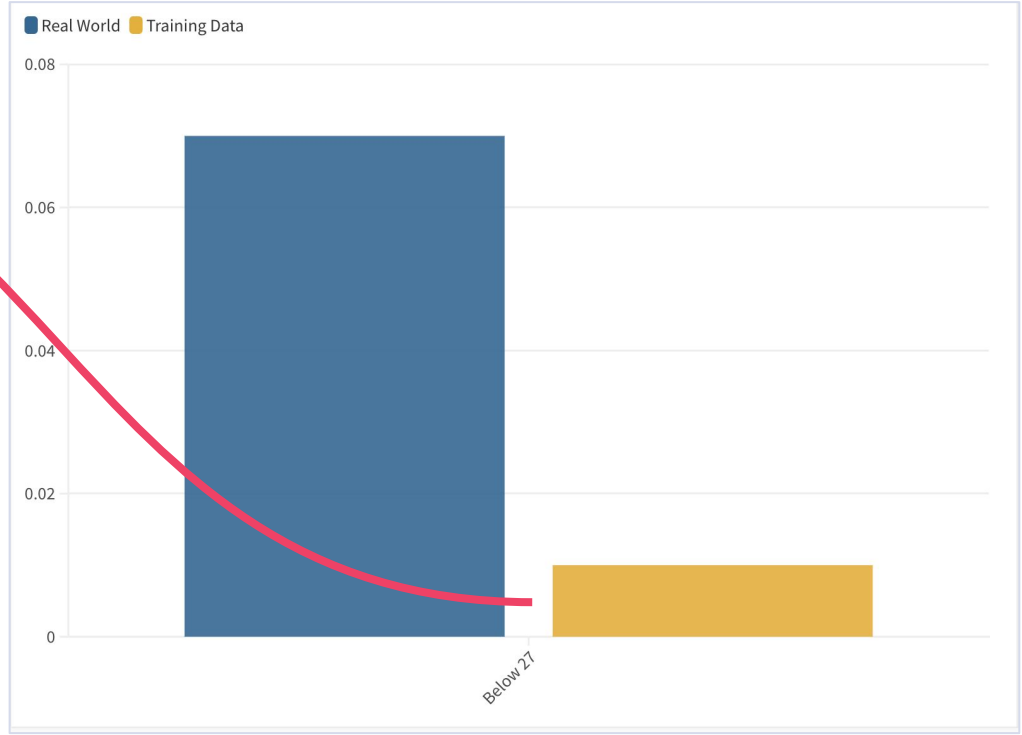
Esta es la distribución por edades de la población real beneficiaria de ayudas sociales en Róterdam, en comparación con los datos de entrenamiento del algoritmo.



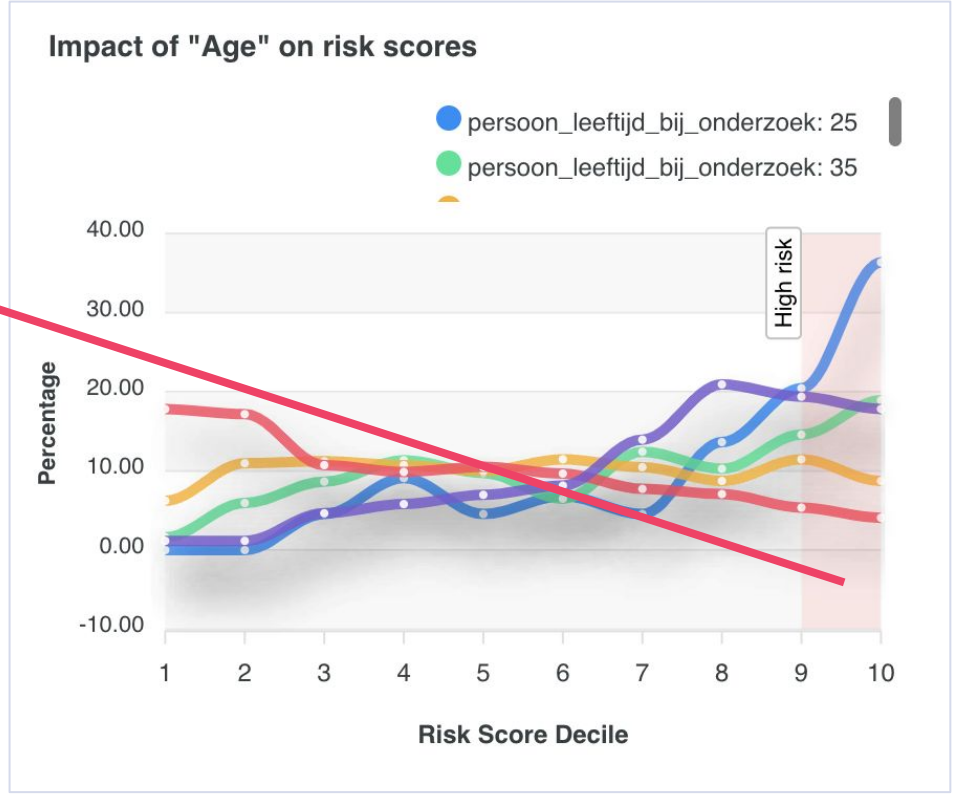
Investiguemos los datos de entrenamiento

¿Notas algo?

Esta es la distribución por edades de la población real beneficiaria de ayudas sociales en Róterdam, en comparación con los datos de entrenamiento del algoritmo.



¿El modelo tiene un sesgo contra los jóvenes?

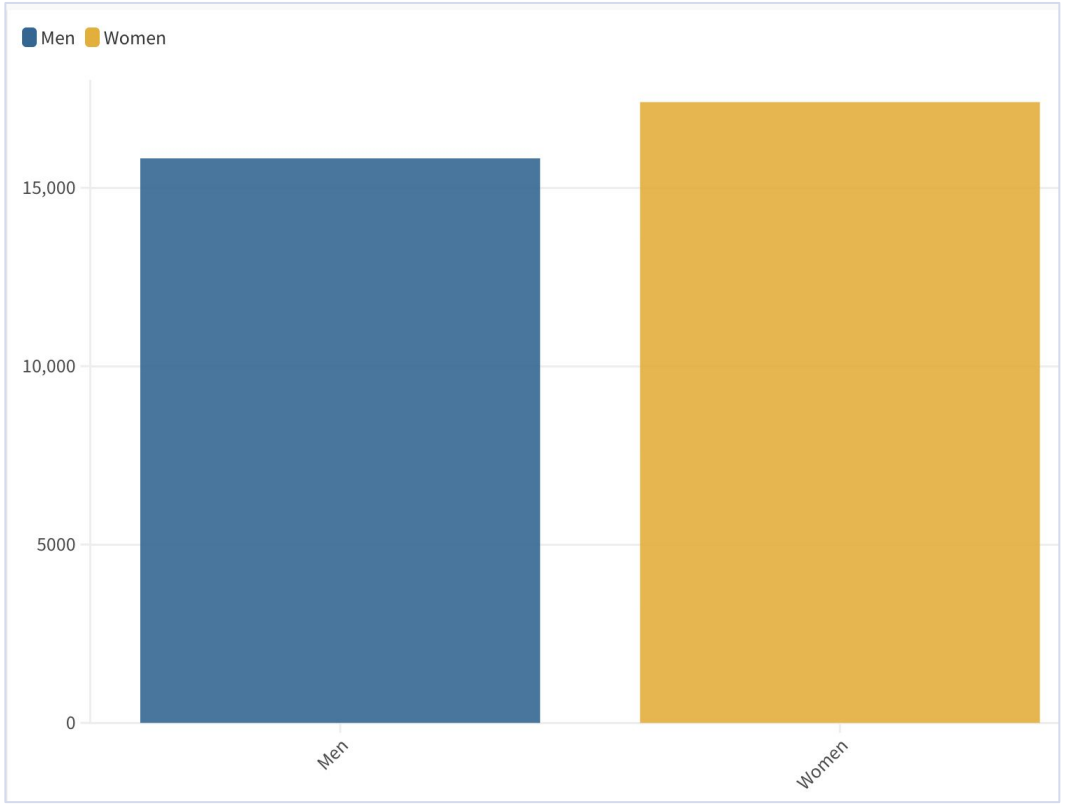


Para el algoritmo, parecía que los jóvenes eran más propensos a cometer fraude en las prestaciones sociales, pero llegó a esa conclusión basándose en **una muestra tan pequeña** que era prácticamente inútil.

¿Tienes tiempo? Comprueba el género.

- Abre https://decision-trees-svelte.vercel.app/
- Busca «género» y comprueba la «distribución de los datos de entrenamiento».
- ¿La distribución por género en los datos de entrenamiento es representativa del mundo real?
- ¿El modelo trata por igual a hombres y mujeres?

Esta es la distribución por género de la población real que recibe asistencia social en Róterdam, en comparación con los datos reales de entrenamiento.



Lecturas adicionales

Informes:

- Machine Bias - ProPublica
- OpenAI GPT clasifica los nombres de los currículos con sesgo racial, según muestra una prueba.
- La IA generativa como Midjourney crea imágenes llenas de estereotipos - Rest of World

Metodologías:

- Metodología de Suspicion Machines - Lighthouse Reports
- Cómo investigamos el sistema de puntuación de personas sin hogar de Los Ángeles – The Markup
- OpenAI GPT clasifica los nombres de los currículos con sesgo racial, según muestra una prueba

Ámbito académico:

- Equidad en el aprendizaje automático: lista de lecturas
- Equidad en el aprendizaje automático: documentación de Fairlearn 0.11.0.dev0
- https://foundation.mozilla.org/en/blog/mozilla-explains-bias-in-ai-training-sets/

AI Spotlight Series

The AI Accountability Network

Pulitzer Center