Instituto Politécnico Nacional
Escuela Superior de Cómputo

De Luna Ocampo Yanina, Ramírez Mendez Kevin, Sainz Takata Juan Pablo Minoru

Estado del Arte

Para empezar a tratar el tema, debemos empezar comprendiendo ¿cuándo y en dónde nace el racismo? para esta pregunta, la respuesta dada es que varios de los especialistas europeos más destacados en el tema sitúan los orígenes del racismo en el nacimiento de la modernidad, primero con la colonización de nuevos territorios y mercados; después, y sobre todo, con la Ilustración, la instalación en las mentes, los corazones y las cartas magnas de Occidente, de la igualdad como valor jurídico central y, finalmente, con la consolidación de los Estados Nación en el siglo XIX.

La discriminación algorítmica se refiere a aquellos procesos a través de los cuales distintos tipos de discriminación que ocurren en el mundo real son reproducidos en entornos de datos o los que surgen exclusivamente en ellos, como cuando los sistemas de reconocimiento facial producen más errores al procesar rostros no caucásicos.

Un ejemplo de esto, son las herramientas que se basan en datos sobre personas, como su edad, género, estado civil, historial de abuso de sustancias y antecedentes penales, para predecir quién tiene una alta probabilidad de estar involucrado en futuras actividades delictivas. Estas herramientas basadas en personas pueden ser utilizadas por la policía, para intervenir antes de que se cometa un delito, o por los tribunales, para determinar durante las audiencias previas al juicio o la sentencia si es probable que alguien que ha sido arrestado reincida, siendo más específicos existe un sistema llamado COMPAS de seguimiento y evaluación policial, se señala que los cálculos algorítmicos para predecir la reincidencia de crímenes violentos alcanzan márgenes de error del 80%. Se mostró que con el uso de esta herramienta, las posibilidades de que las personas afroamericanas sean consideradas con mayor riesgo casi duplica a las de los blancos, aunque nunca reinciden. En cambio, es mucho más probable que las personas blancas sean calificadas de bajo riesgo, aunque posteriormente cometen un nuevo delito. En otras palabras, el problema radica en los datos de los que se alimentan los algoritmos. Por un lado, los algoritmos predictivos son fácilmente sesgados por las tasas de arresto. Según las cifras del Departamento de Justicia de EE. UU., tiene más del doble de probabilidades de ser arrestado si es negro que si es blanco. Una persona negra tiene cinco veces más probabilidades de ser detenida sin causa justificada que una persona blanca. El arresto masivo en Edison Senior High fue solo un ejemplo de un tipo de respuesta policial desproporcionada que no es poco común en las comunidades negras. Cada vez más evidencia sugiere que los prejuicios humanos se han integrado en estas herramientas porque los modelos de aprendizaje automático se entrenan con datos policiales sesgados. Lejos de evitar el racismo, simplemente pueden ser mejores para ocultarlo. Muchos críticos ahora ven estas herramientas como una forma de lavado de tecnología, donde una apariencia de objetividad cubre los mecanismos que perpetúan las desigualdades en la sociedad. El reto consiste en incorporar algoritmos, aprendizaje automatizado y la inteligencia artificial para evitar sesgos humanos, no para reproducirlos. Infortunadamente, los casos de algoritmos los cuales han contribuido a empeorar los procesos, gracias a esto, se han levantado las voces para abogar a favor de la transparencia algorítmica. Un ejemplo de esto, es en Nueva Zelanda, el servicio de inmigración tuvo que cancelar un piloto de modelos predictivos debido a que un medio de comunicación descubrió que el sistema discrimina por nacionalidad. Otro caso, de discriminación algorítmica, fue encontrado en los algoritmos que algunos hospitales utilizan para determinar el nivel de riesgo que una persona enferma tiene, esto mostró que a pacientes negros que estaban más enfermos los ponía en el mismo nivel de riesgo que una persona blanca, este sesgo racial reducía el número de pacientes negros que

necesitaban cuidado extra. Cuando Obermeyer, investigador del aprendizaje automático y gestión de la atención médica en California, y sus colegas realizaron controles estadísticos de rutina en los datos que recibieron de un gran hospital, se sorprendieron al descubrir que a las personas que se auto identificaban como negras generalmente se les asignaban puntajes de riesgo más bajos que a las personas blancas igualmente enfermas. Como resultado de esto, era menos probable que las personas negras fueran remitidas a los programas que brindan una atención más personalizada. Los científicos especulan que este acceso reducido a la atención se debe a los efectos del racismo sistémico, que van desde la desconfianza en el sistema de atención médica hasta la discriminación racial directa por parte de los proveedores de atención médica. Y debido a que el algoritmo asignó a las personas a categorías de alto riesgo en función de los costos, estos sesgos se transmitieron en sus resultados: las personas negras tenían que estar más enfermas que las blancas antes de ser derivadas para recibir ayuda adicional. Solo el 17,7% de los pacientes que el algoritmo asignó para recibir atención adicional eran negros. Los investigadores calculan que la proporción sería del 46,5% si el algoritmo fuera imparcial.
Y aunque pensemos que esto es tán fácil como dejar de considerar la raza como una variable para los cálculos algorítmicos es más complicado de lo que parece. Ya que a la hora de pedir créditos hipotecarios sea cara a cara o por medio de algoritmos estos asignaban tasas de interés más altos a personas negras y nos tenemos que preguntar cómo una máquina puede discriminar sin siquiera poder verte. Entonces también nos tenemos que preguntar qué sesgo existe en los datos para que la máquina pueda discriminar. Y en este caso descubrieron que prestatarios pertenecientes a minorías se les ponían tasas de interés más altas ya que por varias razones estas personas no tienden a comparar tasas de interés, por lo tanto, el algoritmo interpreta esto poniéndoles tasas de interés más altas ya que estos no sabrían que están pagando de más y esto afectaba mayormente a latinos y negros. Estos errores pueden ser involuntarios, pero aún así discriminan a las minorías.
Debido a esto, han empezado a pedir que los organismos manejen algoritmos que realicen un proceso justo y transparente, que expliquen la manera en cómo los sistemas automatizados toman decisiones, especialmente aquellas que afectan significativamente las vidas individuales, como: acceso a prestaciones, un trabajo o a un crédito.
Para terminar con este sesgo racial, dentro de los algoritmos, se ha encontrado que por medio de leyes existentes este se puede mitigar y que los sesgos son más difíciles de corregir si no tenemos diversidad en nuestra sociedad, mayormente en el campo laboral y en políticas públicas.
Para poder combatir esto, como se mencionó previamente, se ha planteado que se debe exigir transparencia no sólo en cuanto al desarrollo de algoritmos, sino también respecto a su uso y quiénes son los beneficiarios de dichos sistemas de toma de decisión, pues son una forma más en la que continúa la vigilancia hacia los usuarios. Esto aunado al hecho que existe una tendencia hacia el monopolio en estos contextos, liderados por Google, Youtube, Facebook. Estas compañías controlan casi en su totalidad la información a la que tenemos acceso y la forma en la que es presentada para nuestro consumo.
Por ejemplo, en 2020 ocurrió el caso de Twitter, donde se hizo evidente que cuando se ponía en una foto la cara de una persona negra y la de una blanca con la distancia suficiente como para que sólo se pudiera mostrar a una de ellas, el algoritmo invariablemente mostraba a la persona blanca. Se movían de lugar las caras, se jugaba con su tamaño, el color de la corbata, el número de personas en la foto, pero el algoritmo impasible priorizaba las caras de personas blancas.

Una vez que se hizo viral el tema, la compañía condujo investigaciones internas al respecto y los resultados mostraron que en efecto existía una preferencia del algoritmo por mostrar a personas

blancas por encima de personas negras. Finalmente ofrecieron una declaración en la que afirmaron que; “Nuestras conclusiones son que no todo en Twitter es un buen candidato para un algoritmo. Y en este caso, cómo cortar una imagen es una decisión hecha mejor por las personas”. Aunque este es un caso muy específico, no es la única compañía que tiene problemas graves con algoritmos racistas; también hay casos ampliamente documentados sobre estas empresas, que muestran discriminación racista hacia creadores, usuarios y los resultados que arrojan determinadas búsquedas en sus páginas.
Debemos incorporar estos hechos a la discusión: ¿Cómo les beneficia a estas compañías incorporar y mantener algoritmos racistas? ¿De qué manera sostienen sus intereses?

Para un campo que supuestamente está remodelando la sociedad como lo son las ciencias de la computación, resulta preocupante observar la homogeneidad de los investigadores dentro del ámbito de la Inteligencia Artificial y es que esta poca diversidad lleva a a la creación de sistemas sesgados que representan de manera pobre a minorías. La inteligencia artificial se ha vuelto una parte fundamental de nuestra vida, pero ¿Qué hacemos si toda esta tecnología masiva está sesgada involuntariamente? y ¿qué hacemos si este campo de estudio es investigado por un sector definido de la población que no la representa?

Si no se tiene diversidad en el grupo de investigación no se podrán reconocer los problemas a los que enfrenta la mayoría de la gente, cuando los problemas no nos afectan, no pensamos en ellos, no les damos la importancia o inclusive no somos conscientes de su acontecer, porque no interactuamos con personas que los experimenten.

¿Hay alguna forma de atacar estos sesgos?
La principal razón por la que necesitamos diversidad, tanto en los conjuntos de datos como en los grupos de investigadores es porque se necesita de personas que tengan el sentido social de cómo son las cosas. Desde un punto de vista técnico, hay distintos acercamientos para reducir los sesgos. Uno es diversificar el conjunto de datos y tener anotaciones sobre datos relevantes del mismo, una vez entrenado el modelo, probar que tan bien trabaja con diferentes subgrupos. Incluso siguiendo esto, no podemos asegurar que no exista un sesgo, puesto que no se puede tener un conjunto de datos que defina a la perfección a todas las personas y un modelo que no generaliza, no es un buen modelo.

Estos sistemas de machine learning están siendo utilizados para tomar decisiones que cambian vidas. Estas decisiones pueden llegar a perjudicar los derechos humanos, usualmente de las personas más vulnerables de la sociedad.

Bajo un buen uso y diseño, los sistemas de machine learning, estos pueden ayudar a eliminar algunos sesgos en la toma de decisiones que afectan a la sociedad. Sin embargo, también es posible para estos sistemas reforzar el sesgo sistemático y la discriminacion, eludiendo la garantía de dignidad humana. Como ejemplo de estos casos tenemos:

- Al traducir un texto frases referentes a mujeres normalmente se traducen a un sujeto masculino.
- Cámaras que tienden a interpretar a las personas asiáticas como que siempre están parpadeando

- Las Técnicas de Word Embedding de Procesamiento de Lenguaje Natural que tienden a clasificar nombres Euroamericanos como agradables y afroamericanos como desagradables
- En medicina, investigadores usaron técnicas de aprendizaje profundo para determinar el cancer de piel, el conjunto de datos utilizado contenía solamente un 5% de imagenes de individuos de tez oscura y el algoritmo no fue probado en personas de color. Es por ello que el desempeño del algoritmo varía dependiendo de la población.

Los resultados discriminatorios no sólo violan los derechos humanos, sino que también demeritan la confianza pública en los algoritmos de machine learning, dificultando el desarrollo del campo y mitigando su potencial tanto social como económico.

Es por ello que se necesitan soluciones sistemáticas para atacar el problema de discriminación algoritmica.

Usualmente, no se pone mucha atención al cómo se recolectan, procesan y organizan los datos, llevando así a que algunos grupos sean representados en gran medida y otros son infrarepresentados.
Ejemplo: Más del 45% de los datos de ImageNet provienen de los Estados Unidos, siendo una muestra no representativa al significar el 4% de la población mundial. En contraste, China e India representan un 3% de los datos de ImageNet, cuando estos países representan el 36% de la población mundial.

También se debe considerar que los sesgos en los datos usualmente reflejan desbalances en la sociedad.
Ejemplo: Wikipedia parece una fuente diversa enriquecida de datos, pero menos del 18% de las entradas biográficas son de mujeres.

Algunas medidas que se han puesto para desvanecer la discriminación algoritmica han sido:

Un grupo de investigadores formando el proyecto data nutrition, el cual busca crear herramientas y prácticas que alienten al desarrollo responsable de IAs. Este proyecto sugiere agregar metadatos a nuestro conjunto de datos que describen la calidad de los mismos.

Otro enfoque complementario es el uso de las mismas técnicas de machine learning para identificar y cuantificar sesgos en algoritmos y en datos. A este modus se le conoce como IA de auditoría, en la cual el auditor es una algoritmo que sistemáticamente prueba el modelo de machine learning original para identificar sesgos en ambos el modelo y los datos de entrenamiento. Un ejemplo de este enfoque se puede observar al usar la técnica word embedding para cuantificar estereotipos históricos. Esta técnica mapea cada palabra a un punto en el espacio, tal que la distancia entre vectores captura similitudes semánticas entre palabras. Capturando de esta forma analogías, tales como *hombre es a rey como mujer es a reina*. En este caso se usó una IA de auditoría para consultar a la técnica de word embedding por otras analogías de género. Revelando de esta forma analogías tales como *hombre es a doctor como mujer es a enfermera* o *hombre es a programador como mujer es a ama de casa*. Una vez que el auditor revela estos estereotipos en el modelo principal es posible reducir este sesgo modificando el mapeo de los vectores de palabras.

Instituto Politécnico Nacional
Escuela Superior de Cómputo

De Luna Ocampo Yanina, Ramírez Mendez Kevin, Sainz Takata Juan Pablo Minoru

Título
Elaboración de un protocolo que disminuya la discriminación racial algorítmica en la creación de algortimos de machine learning a partir del análisis de algoritmos discriminatorios implementados durante el periodo de 2016 - 2021 para su adaptación en México.

Planteamiento del problema
¿Cómo evitar la discriminación racial algorítmica en la creación de algortimos de machine learning a partir del análisis de algoritmos discriminatorios implementados durante el periodo de 2016 a 2021 con la creación de un protocolo para su adaptación en México?

Preguntas
¿Qué es la discriminación?
¿Qué es el racismo?
¿Qué es machine learning?
¿Qué es la discriminación racial algorítmica?
¿Qué es un protocolo?

Objetivo general
- Elaborar un protocolo que disminuya la discriminación racial algorítmica en los futuros algoritmos de machine learning adaptados a la sociedad mexicana a partir del análisis de algoritmos discriminatorios implementados durante el periodo 2016 - 2021.

Objetivos específicos
- Encontrar las principales razones por las cuales ocurre el sesgo racial dentro de los algoritmos discriminantes.
- Investigar cómo el racismo afecta a los individuos de forma socioeconómica.
- Entender cómo funcionan las técnicas utilizadas de machine learning
- Observar cómo están entrenados los algoritmos para que estos realicen. la discriminación racial algorítmica
- Adaptar, dentro de México, las soluciones que se han propuesto con respecto a los algoritmos discriminatorios.

Variables
Dependiente = Protocolo que disminuya la discriminación racial algorítmica
Independiente = Análisis de algoritmos discriminatorios

Hipótesis
Con la elaboración de este protocolo se quiere compartir en México el conocimiento que existe de cómo los algoritmos discriminatorios impactan de forma negativa a diversos aspectos de la sociedad donde se implementan y el cómo estos pueden afectar a los mexicanos, apoyando la teoría de que los algoritmos discriminatorios llegan a serlo por un mal uso de datos ya que los algoritmos aprenden de los datos ya existentes.

Justificación

Instituto Politécnico Nacional
Escuela Superior de Cómputo

De Luna Ocampo Yanina, Ramírez Mendez Kevin, Sainz Takata Juan Pablo Minoru

La presente investigación se enfocará en el estudio de algoritmos discriminantes, analizando el por qué estos discriminan, ya que el uso de ellos afecta entornos sociales, laborales y políticos y puede llegar a agrandar las brechas de desigualdad ya existentes. Por lo tanto, se busca analizar la causa por la cual estos algoritmos discriminan y con este análisis proponer mejores prácticas tomando como centro del estudio a los mexicanos, para que ya no se vean tan afectados por la discriminación presente en el uso de servicios que pueden utilizar este tipo de algoritmos.

Se pretende ahondar en esta problemática mediante el planteamiento de las bases para un proceso de creación de algoritmos de machine learning donde se busque disminuir la discriminacion algorítmica mediante la elaboración de un protocolo que ahonde en los procesos para la elaboración de algoritmos no discriminatorios de Machine Learning.

Esta investigación tiene como fin llenar el hueco de conocimiento que existe en México sobre este tipo de algoritmos y el cómo estos pueden afectar a los mexicanos, ya que no existe un estudio sobre México y no se puede generalizar hallazgos sobre el tema hechos en otras partes del mundo porque contamos con diferentes entornos socioeconómicos.

Los resultados que esperamos obtener pueden apoyar a la teoría de que los algoritmos discriminatorios discriminan por un mal uso de datos ya que los algoritmos aprenden de los datos existentes.

Cronograma de actividades

Fecha / Actividades	Agosto				Septiembre				Octubre				Noviembre				Diciembre				Explicacción
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1. Estado del Arte																					Investigamos acerca del contexto para poder acotar nuestro tema de investigación.
2. Planteamiento del problema																					El tema selecto que tenemos como objetivo contestar.
3. Preguntas de investigación																					Planteamos preguntas para tener en mente lo que debemos contestar dentro de nuestr
4. Objetivos																					Expectativas y resultados que se esperan obtener en el estudio.
5. Justificación																					Explicar la importancia de nuestro tema de investigación.
6. Marco teórico y metodológico																					El fundamento de nuestra investigación.
7. Desarrollo de la investigación																					Abordamos y desglosamos el tema que propusimos.
8. Conclusión. Bibliografiias / Referencias																					Concluímos nuestra investigación y nuestro material de apoyo.
9. Entrega del artículo																					Enviar el artículo para revisión.
10. Exposición																					Exponer nuestro tema.

De Luna Ocampo Yanina, Ramírez Mendez Kevin, Sainz Takata Juan Pablo Minoru

Bibliografías:

1. Turner Lee, N. (2018), "Detecting racial bias in algorithms and machine learning", Journal of Information, Communication and Ethics in Society, Vol. 16 No. 3, pp. 252-260. https://doi.org/10.1108/JICES-06-2018-0056

2. Snow, Jackie. (2018) ""We're in a diversity crisis": cofounder of Black in AI on what's poisoning algorithms in our lives". *MIT Technology Review*.

3. Kochi Erica "White Paper: How to Prevent Discriminatory Outcomes in Machine Learning". *World Economic Forum*. 12 March 2018.

4. Zou, James; Schiebinger, Londa (July 2018). "AI can be sexist and racist — it's time to make it fair". *Nature*. **559** (7714): 324–326. Bibcode:2018Natur.559..324Z. doi:10.1038/d41586-018-05707-8. PMID 30018439.

5. Simonite, Tom (2018) "AI Is the Future—But Where Are the Women?". *Wired*. ISSN 1059-1028.

6. Obermeyer, Powers, Vogeli, & Mullainathan. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. 10.1126/science.aax2342

7. Köchling, A., Wehner, M.C. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus Res 13, 795–848 (2020). https://doi.org/10.1007/s40685-020-00134-w

8. How Some Algorithm Lending Programs Discriminate Against Minorities. (2018, 24 noviembre). Npr. Recuperado 24 de agosto de 2022, de https://choice.npr.org/index.html?origin=https://www.npr.org/2018/11/24/670513608/how-some-algorithm-lending-programs-discriminate-against-minorities

9. Cave, S., Dihal, K. The Whiteness of AI. *Philos. Technol.* 33, 685–703 (2020). https://doi.org/10.1007/s13347-020-00415-6

10. Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings*, 111: 49-54.

11. Olivia Gall, Ermanno Vitale, Sylvia Schmelkes. 2005. “La Discriminación Racial” pdf.

12. Obermeyer , Z. , Powers , B. , Vogeli , C. Mullainathan , S. *Science* 336 , 447–453 (2019).

13. Vlasceanu, M. , Amodio, D. (2022) “Propagation of Societal Gender inequality by internet search algorithms”, National Academy of Sciences, PubMed ID 35858360, ISSN 00278424

14. W. D. Heaven. "Predictive policing algorithms are racist. They need to be dismantled". MIT Technology Review.

    https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/ (accedido el 25 de agosto de 2022).

15. C. Buenadicha, G. Galdon Clavell, C. Pombo, D. Loewe y M. Paz Hermosilla, *LA GESTIÓN ÉTICA DE LOS DATOS*. BID, 2019.

16. M. Rateike, O. Mineeva, A. Majumdar, K. Gummadi, I. Valera, (2022) “Don't Throw it Away! The utility of unlabeled Data in fair Decision Making”

17. J. Ogata. "Coolhuntermx - Tecnología y discriminación, racismo en los algoritmos". Coolhuntermx. https://coolhuntermx.com/racismo-en-los-algoritmos-mundo-digital-jumko-ogata/ (accedido el 25 de agosto de 2022).

18. Ángel Jiménez de Luis. "¿Es racista el algoritmo de Twitter?" ELMUNDO. https://www.elmundo.es/tecnologia/2020/09/22/5f68efe3fc6c83b9088b465b.html (accedido el 25 de agosto de 2022).

19. P. Rivas Vallejo, *Discriminación algorítmica: detección, prevención y tutela*. Madrid, 2021.

20. A. Pandey y A. Caliskan, *Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms*. 2021.