La era de los algoritmos. ¿Enemigos o aliados?

Introducción a la IA .Algoritmos. Machine learning

Para hablar de Ia, Algoritmos y machine learning previamente tenemos que mencionar un elemento fundamental. Los datos. La relación entre la IA y los datos es muy directa

¿Qué es un dato?

Tipos de datos:

Datos personales "genéricos". Los datos personales son definidos como la información de cualquier tipo referida a personas determinadas o determinables. Todo aquello que hace que seas Juan y no Pedro.

Datos sensibles. Los datos "sensibles" conforman una categoría especial. Son aquellos que afectan la esfera íntima del titular y que pueden generar discriminación arbitraria. A modo de ejemplo, las normas mencionan, dentro de esta categoría, a aquellos que revelan origen racial, étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, participación o afiliación en una organización sindical o política, información referida a la salud, preferencia o vida sexual.

Datos biométricos Son aquellos obtenidos a partir de un tratamiento técnico específico, relativos a las características físicas, fisiológicas o conductuales de una persona humana, que permitan o confirmen su identidad única..

Datos genéticos. Son aquellos relativos a las características genéticas heredadas o adquiridas de una persona humana que proporcionen información sobre su físico o salud, obtenido mediante un análisis de muestra biológica. Esta categoría no existe en la ley argentina, pero sí se incluye de manera diferenciada en el Proyecto.

Datos de acceso irrestricto. Existen ciertos datos cuyo tratamiento no requiere consentimiento. En algunas de las normas analizadas, se alude a los siguientes: los que se limiten al nombre y apellido, documento nacional de identidad, identificación tributaria y previsional, ocupación, fecha de nacimiento, domicilio, correo electrónico, así como aquellos necesarios para el tratamiento de la información crediticia [(24)].

TIPOS DE TRATAMIENTO DE DATOS PERSONALES

Los datos personales son sometidos a tratamientos.

Tratamiento. Las normas, estándares y recomendaciones se refieren al tratamiento

en sentido estricto [(26)], o a cualquier operación o procedimiento organizado, electrónico o no, que permita la recolección, conservación, ordenación, almacenamiento, modificación, relacionamiento, evaluación, bloqueo o destrucción y todo procesamiento de datos personales en general. También abarca la cesión de datos personales a través de comunicaciones, consultas, interconexiones o transferencias.

Tratamiento electrónico. El tratamiento electrónico o digital que mencionan las normas se vincula con la utilización de tecnologías de la información y comunicación (en adelante, TIC). Aunque las normas en general no diferencian ambos tratamientos en las definiciones, es útil distinguir esta categoría, porque será esencial para comprender la diferencia con el tratamiento automatizado. En el tratamiento electrónico, siempre existe un operador humano que, a través del uso de TIC, efectúa la recolección, conservación, ordenación, almacenamiento, modificación, relacionamiento, evaluación, bloqueo o destrucción de los datos tratados. **Tratamiento automatizado**. El tratamiento automatizado de datos se encuentra regulado en el Convenio 108 del Consejo de Europa y en el Reglamento 679/2016 de la Unión Europea. También mencionan este tipo de tratamiento la Ley de Costa Rica, Chile, Perú, Panamá, Uruguay, los Estándares de la Red Iberoamericana y la Ley de Brasil **La diferencia entre el tratamiento automatizado y el tratamiento electrónico se relaciona con la utilización de técnicas de inteligencia artificial.** Especialmente, una de las más sofisticadas, que se conoce como aprendizaje profundo *(deep learning)*, basado en redes neuronales complejas (ampliar estas cuestiones, en *infra* punto II).

¿Qué es la IA? Son sistemas de software (y en algunos casos también de hardware) diseñados por seres humanos que, dado u objetivo complejo, actúan en la dimensión física o digital mediante la percepción del entorno a través de la obtención de datos, la interpretación de los datos estructurados o no estructurados que recopilan (...) y decidiendo la acción o acciones óptimas que deben llevar a cabo para lograr el objetivo establecido. Los sistemas de IA pueden (...) adaptar su conducta mediante el análisis del modo en que el entorno se ve afectado por sus acciones anteriores

Definición de principales Capacidades y Disciplinas científicas", Grupo independiente de expertos de alto nivel sobre inteligencia artificial, abril 2019.

¿Qué es un algoritmo? Conjunto de instrucciones o reglas definidas , ordenadas que permite solucionar un problema, realizar un cómputo, procesar datos y llevar a cabo tareas o actividades, siguiendo pasos sucesivos, llegando a un resultado final. Los algoritmos son la base de la IA, que ejecutan instrucciones a partir de diversas técnicas, para transformar datos en patrones de información, luego en conocimiento y, desde allí, automatizar tareas, elaborar predicciones o previsiones.

Machine learning: El término *aprendizaje automático* se refiere a la detección automatizada de patrones significativos en los datos. En las últimas dos décadas se ha convertido en una herramienta común en casi cualquier tarea que requiera la extracción de información de grandes conjuntos de datos. Existen diversas técnicas de machine learning, como aprendizaje supervisado, no supervisado, aprendizaje por refuerzo, y también las redes neuronales profundas o deep learning.
Los modelos de machine learning son entrenados, cargados de una gran cantidad de datos.Es por ello que para llegar a hablar de sesgos tenemos antes que hablar de datos, y su tratamiento.
Entonces, ¿Cómo relacionamos estos conceptos?
Tenemos→ Grandes cantidades de datos→ estos son procesados con ALGORITMOS de Machine learning→ como resultado se establecen **patrones.**

Ahora bien, una vez dada esta introducción podemos pasar a nuestro tema principal, que son los Sesgos, y sus posibles soluciones al realizar tratamiento de datos.

SESGOS
¿Que es un sesgo? En palabras simples, es omitir considerar información relevante.
Sesgos algorítmicos: Un sistema informático refleja los valores de los humanos que están implicados en la codificación y recolección de datos usados para entrenar un algoritmo.

FUENTE BASE: https://users.dcc.uchile.cl/~rbaeza/bias/sesgos-algoritmos.html
RICARDO BAEZA

Otras definiciones:

Del inglés BIAS: “La acción de apoyar u oponerse a una persona o cosa en particular de manera injusta, debido a que permite que las opiniones personales influyan en su juicio” (Cambridge English Dictionary).

SESGO: Error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras” (RAE)

DÓNDE PUEDE PRESENTARSE EL SESGO:

- **EN LOS DATOS:**

  MUESTRA (CANTIDAD) – CALIDAD

  INCOMPLETOS, INCORRECTOS, INSUFICIENTES

  EL DATO MISMO PUEDE CONTENER UN PREJUICIO - LENGUAJE CONTENIDO EN LOS DATO

- EN EL ALGORITMO

DISEÑO – FUNCIÓN DE ÉXITO (resultado)

SOBRE QUÉ SON LOS SESGOS

→ 3 TIPOS DE SESGOS CLÁSICOS:

- **ESTADÍSTICO:** Cómo obtenemos los datos, de errores de medidas o similares. **EJEMPLO: si la policía está presente en algunos barrios más que en otros, no será extraño que la tasa de criminalidad sea más alta donde tenga mayor presencia (o en otras palabras, mediremos más donde está uno de los instrumentos de medida).**

- **CULTURAL:** Aquel que deriva de la sociedad, del lenguaje que hablamos o de todo lo que hemos aprendido a lo largo de la vida. **Ejemplo**: **Los estereotipos de las personas de un país son un ejemplo claro**

- **COGNITIVO:** Aquel que nos identifica y que depende de nuestra personalidad, de nuestros gustos y miedos. **Ejemplo si leemos una noticia que está alineada con lo que pensamos, nuestra tendencia será validarla aunque sea falsa.**

→ DE ESTOS 3 SESGOS (PRINCIPALES) DERIVAN OTROS MÁS:

SESGO DE GÉNERO: No hay suficiente incorporación de la teoría lingüística feminista en los procesos de ML. Esto se traduce en sesgos en la forma de

nombrar; el orden de preferencia; las descripciones sesgadas; el uso y la tipología de metáforas y el grado de presencia/ausencia de las mujeres en los textos escritos. (Susan Leavy, University college dublin)

*

https://www.eldiario.es/andalucia/desdeelsur/sesgos-genero-lenguaje-inteligencia-artificial_132_1171463.html

SESGO DE CONFIRMACIÓN: La tendencia a favorecer, buscar, interpretar, y recordar, la información que confirma las propias creencias o hipótesis, dando desproporcionadamente menos consideración a posibles alternativas

SESGO DE ORDEN (RANKING): Se da cuando buscamos en la web, ya que las personas tienden a hacer clics en las primeras posiciones y el buscador podría interpretar que estas respuestas son mejores que las siguientes.

SESGO DE PRESENTACIÓN: Se encuentra en las recomendaciones en el ámbito del comercio electrónico. Solo aquello que se muestra al usuario podrá tener clics. Todo lo que no salga en la página de resultados no puede ser escogido.

Esto puede parecer obvio, pero la verdad es que no todos dan cuenta de ello. La única forma de romper con el ciclo es que se muestra el universo total de resultados (materialmente imposible)

El sesgo de presentación tiene relación con el **FILTRO BURBUJA**: El sistema muestra únicamente aquello que te gusta. Como se basa en las acciones del pasado, no se puede ver lo que se desconoce.

SESGOS DE SEGUNDO ORDEN: Por ejemplo, cuando una persona usa la información de los primeros resultados de un buscador y la reutiliza para escribir nuevos artículos. Esto significa que cuando la información es recolectada, posiblemente ya esté sesgada y el buscador crea que son aún más relevantes.

*Existen muchísimos más (investigaciones han identificado al menos un centenar de sesgos).

→ **Ahora bien, existe una relación de dependencia entre los datos que alimentan al algoritmo, y el algoritmo mismo.** En este contexto los sesgos que un modelo puede adquirir en relación con los datos con los que es entrenado, serían al menos los 3 siguientes:

(Fuente: Sesgo e Inferencia en redes neuronales ante el derecho. Carlos Amunategui y otros)

Sesgos algorítmicos: Un sistema informático refleja los valores de los humanos que están implicados en la codificación y recolección de datos usados para entrenar un algoritmo.

Sesgo algoritmico:

SESGO DE INTERACCIÓN: El propio usuario o programador introduce de forma inadvertida un sesgo en el modelo por la manera en que interactúa con él.

SESGO LATENTE: Cuando el modelo realiza correlaciones inapropiadas, generalmente al establecer falsos nexos entre puntos de datos.

SESGO DE SELECCIÓN: Cuando la base de datos no es suficientemente representativa de la diversidad existente en el medio social.

Importante:

EL SESGO EN EL APRENDIZAJE AUTOMÁTICO SE PUEDE DETECTAR Y DISMINUIR CON BASTANTE FACILIDAD VS. SESGOS EN HUMANOS

NECESIDAD DE INVERTIR EN HERRAMIENTAS DE EVALUACIÓN, CONTROL Y MITIGACIÓN.

Al decir que los algoritmos presentan sesgos, estamos trasladando la responsabilidad a los sistemas de inteligencia artificial, cuando en realidad, son las personas tras los sistemas, quienes deben preocuparse y en consecuencia, ocuparse de evaluar, controlar y mitigar resultados que pueden potencialmente generar daños.

CASOS CONTROVERSIALES: DISCRIMINACIÓN ALGORÍTMICA

Si tomamos por un lado lo que es la IA explicado inicialmente y la existencia de múltiples sesgos (cognitivos y de IA) por otro, obtenemos como resultado, el riesgo (existente) de que sistemas basados en ML tomen decisiones que reproduzcan e intensifiquen discriminaciones y en consecuencia, se perpetúe una sociedad en que abundan las injusticias.

Basta una rápida búsqueda en Google, introducir las palabras “IA” y “Sesgos” para pensar que la situación se está saliendo de control. Hay muchísimos casos controversiales y la repercusión y el escándalo a su alrededor es en cierto modo legítima, por cuanto se han vulnerado personas y sus derechos, y han sido decisiones que han cambiado vidas, afectando por lo general, a grupos más vulnerables.

Por mencionar algunos casos:

Caso Amazon – 2014: Agente de selección de personal. En 2014, la empresa tecnológica quiso automatizar su selección de currículos a fin de confeccionar una lista corta de candidatos a un puesto determinado. Para ello construyó un algoritmo que se basaba en la información de los empleados que la firma ya tenía, su tasa de retención y ascensos. Puesto que la mayor parte de los empleados de Amazon eran hombres, el algoritmo procedió a eliminar de la lista de candidatos a las mujeres, construyendo listas cortas exclusivamente masculinas. Lo más inquietante es que en los currículos no se incluía la mención del sexo del solicitante, pero, aparentemente, el agente lo infirió de otros datos incluidos en el currículo. Como el algoritmo no pudo ser corregido para lograr la objetividad inicialmente deseada, la compañía debió retirarlo.

Caso Google – 2015: La aplicación de reconocimiento facial Google Photos, etiquetó una de las fotos de Jacky Alcine, ciudadana afroamericana, con la palabra "gorila".

Esto sucedió porque el algoritmo no había sido entrenado con suficientes imágenes de personas de piel oscura.

Caso Microsoft – 2016: "Tay", era un *chatbot* cuyo fin era imitar el comportamiento de una adolescente curiosa y entablar en las redes sociales una conversación informal y divertida con una audiencia de entre 18 y 24 años. El proyecto mostraría las promesas y el potencial de las interfaces conversacionales alimentadas por inteligencia artificial. Sin embargo, en menos de 24 horas, el "inocente" Tay a través de *tweets*, mostraba su empatía hacia Hitler o su apoyo al genocidio al responder a preguntas de los usuarios de las redes sociales, son algunos ejemplos, además de insultos raciales y comentarios sexistas y homófobos. También defendió el Holocausto, los campos de concentración o la supremacía blanca, y se mostró contraria al feminismo.

Caso Universidades UK – 2020: Con la pandemia del Covid-19, las universidades de UK optaron por buscar una alternativa a los exámenes presenciales de admisión.

Estudios sobre los puntajes anteriores habían demostrado la existencia de sesgos en función de la edad, género y etnia, siendo entonces la equidad una cuestión de especial importancia a combatir. Teniendo eso en mente, las autoridades resolvieron utilizar un algoritmo.

Casi el 40% de los estudiantes terminaron recibiendo puntajes de exámenes rebajados de las predicciones de sus maestros, lo que amenazó con costarles sus lugares universitarios. El análisis del algoritmo también reveló que había dañado de manera desproporcionada a los estudiantes de la clase trabajadora y comunidades

desfavorecidas e inflado las puntuaciones de los estudiantes de las escuelas privadas.

Finalmente, se revocó la decisión de utilizar el algoritmo como elemento decidor de los procesos de admisión y los estudiantes ahora recibirán las puntuaciones previstas por su profesor o por el algoritmo, la que sea más alta.

Así entonces, no pretendemos desconocer que los algoritmos utilizados en el mercado laboral, bancos, compañías de seguros, entidades comerciales, sistemas educativos, etc. estén provistos de determinados valores e ideologías en su ADN. Sin embargo, queremos centrar nuestra presentación en las posibles soluciones que se están desarrollando hoy a estos problemas.

Ciertamente no se le puede exigir objetividad y justicia a las entidades comerciales, a los bancos o a las compañías de seguro. Hay en ellos una intención de lucrar y de obtener la mayor ganancia posible a cualquier costo. Distinto es el caso de las redes de empleabilidad, de los sistemas de educación, de ayuda social y de justicia. En ellos debe existir una garantía de equidad.

SESGO HUMANO VS. SESGO ALGORÍTMICO: S. MULLAINATHAN

Es importante abrir los ojos y entender que los sesgos no son creados por la IA, y que no podrán simplemente desaparecer. Se requiere un cambio más profundo a nivel de estructuras, sistemas y de sociedad.

A pesar de lo catastrófico que han resultado casos como los mencionados y otros tanto más, pareciera que los sistemas de IA cuentan con la virtud de ser más flexibles y aprender mejor y más rápido de sus errores, en comparación con nosotros los humanos. Los sesgos en la IA serían más fáciles de corregir, que los sesgos que tenemos nosotros los humanos.

El año 2004 el Marianne Bertrand y Sendhil Mullainathan publicaron un estudio llamado: *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination.* En Diciembre del 2019 ese estudio adquirió nueva relevancia.

La primera publicación (2004) dio cuenta de un experimento en el cual se midió la discriminación racial en el campo laboral. Se respondió a diversas ofertas de trabajo con CVs ficticios. A cada CV se le asignó aleatoriamente un nombre que sonaba “muy afroamericano” o “muy blanco” . Los resultados mostraron una discrminación significativa contra los nombres afroamericanos, obtenías menos entrevistas de trabajo.

El año 2019 Mullainathan realizó otro experimento, en el cual también se buscaba medir la discriminación racial. Dos pacientes buscaban atención médica. Ambos

estaban lidiando con la diabetes y la presión arterial alta. Un paciente era negro, el otro era blanco. El primer paciente recibió una atención peor.

Las conclusiones son las mismas, sin embargo existe una gran diferencia entre una investigación y otra, el año 2004 fueron personas encargadas de los procesos de contratación quienes tomaron las decisiones. El año 2019, fue un programa de computadora quien asignó determinado puntaje para acceder a servicio de salud.

Entre uno y otro estudio puede observarse la existencia de sesgos, humanos por un lado, del algoritmo por otro. Pero otra gran diferencia es lo que se requiere para develar y corregir la existencia de ese sesgo.

Identificar el comportamiento discriminatorio de un grupo particular de personas (gerentes de contratación) suele ser muy difícil. Por el contrario, descubrir la discriminación algorítmica fue mucho más sencillo. "*Este fue un ejercicio estadístico, el equivalente a preguntar al algoritmo "¿qué harías con este paciente?" cientos de miles de veces y trazando las diferencias raciales. El trabajo era técnico y rutinario, y no requería ni sigilo ni ingenio". (Cita: Artículo NY: https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html)*

Los seres humanos son inescrutables de una manera que los algoritmos no lo son. Nuestras explicaciones de nuestro comportamiento están cambiando y construidas después de los hechos. Para medir la discriminación racial por parte de las personas, debemos crear circunstancias controladas en el mundo real donde solo la raza difiere. Para un algoritmo, podemos crear igualmente controlados simplemente alimentándolo con los datos correctos y observando su comportamiento. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

¿Es la inteligencia artificial un enemigo?¿Cómo podemos hacer de la inteligencia artificial nuestro aliado?

La **red iberoamericana de protección de datos personales**, en la redacción de las Orientaciones especificas para el cumplimiento de los principios rectores de la protección de datos personales establece:

-Que el modelo de la IA no debe enfatizar la información relacionada con el origen racial o étnico, opinión política, religión, orientación sexual, y además deja sin dudas que se debe ESTABLECER un sistema de monitoreo constante del modelo de IA con la finalidad de identificar la existencia de sesgos y en la medida de lo posible implementar una gestión de riesgos, obteniendo como producto final reportes y estadisticas que permitan analizar los resultados. (datos sensibles)

Los datos de entrada estan afectados por dos variables en principio: los sesgos (incorporación de datos parciales, insuficientes, no actualizados o manipulados) y la pertinencia (relevancia, inconsistencia o completitud de datos). El desarrollo del algoritmo, se puede ver afectado por los patrones (sesgos de la lógica de programación, inclusión de funciones utilizadas para su codificación) y los errores (condiciones de la operación que reflejan un funcionamiento diferente al previsto y atentan contra las premisas del diseño. Por último los riesgos están relacionados con la pertinencia y precisión de los resultados del algoritmo y como respuesta al análisis de datos de entrada.

POSIBLES SOLUCIONES

Sugerimos dividir las posibles soluciones en **ÁREA TECNICA** (es decir soluciones para el sistema mismo) y **HABILIDADES HUMANAS O EDUCACIÓN ÉTICA** (teniendo en cuenta que los sistemas son realizados por humanos, al menos por el momento, y que factores como la calidad del dato recolectado, es decir las características pueden ser menos informativas o recopiladas de manera menos confiable para ciertas partes de la población, los sesgos que tenemos los seres humanos como inherentes a la existencia misma, influyen de manera directa en los sistemas que se desarrollan)

-AREA TÉCNICA:

¿Cómo aprenden a discriminar los modelos?
https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

-muestra sesgada
-ejemplos contaminados

-características limitadas
-disparidad del tamaño de la muestra

Un conjunto de características que admite precisión para el grupo mayoritario puede no serlo para un grupo minoritario.
Diferentes modelos con la misma precisión informada pueden tener una distribución muy diferente entre la población.
¿Cómo podemos solucionarlo?

Modelo Fair ML

Fairlearn es un nuevo paquete de Python desarrollado por Microsoft. Implementa varios algoritmos para detectar y mitigar problemas de equidad de grupo en modelos de aprendizaje automático.
¿Cómo funciona?
Evaluación de la equidad: Fairlearn contiene un Fairlearn Dashboard componente y un conjunto de métricas que lo ayudan a medir la equidad de su modelo.

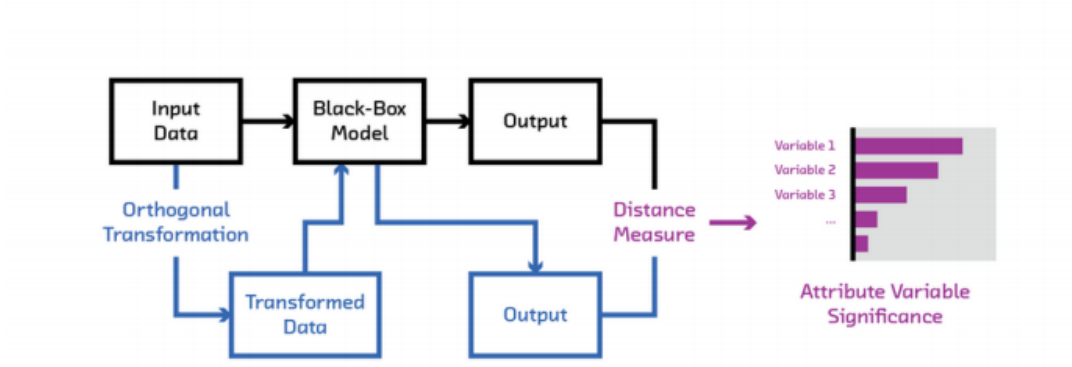
Mitigar la injusticia: junto con las métricas y el tablero, hay un conjunto de algoritmos fairlearn para ayudar a mitigar el comportamiento injusto en los modelos. Puede, por ejemplo, utilizar un algoritmo de posprocesamiento para mejorar los modelos existentes. Pero también hay un par de algoritmos que te ayudan a mejorar el modelo durante el entrenamiento.
El paquete fairlearn contiene un componente llamado FairlearnDashboard. Es un widget que podemos usar dentro de un cuaderno de Python que mide y visualiza dos métricas para nuestro modelo:

¿Cómo puedo utilizar fairlearn para mejorar un modelo injusto?
El paquete fairlearn contiene varios algoritmos que ayudan a resolver la injusticia en los modelos sin cambiar los datos que usamos para entrenar el modelo.

Hay dos estrategias que podemos aplicar para resolver la injusticia con fairlearn:

Para los modelos existentes, podemos mitigar la injusticia mediante el posprocesamiento.
Para los nuevos modelos, podemos utilizar algoritmos de reducción para mejorar la equidad.

Usar el algoritmo ThresholdOptimizer para mejorar un modelo existente
TresholdOptimizer se basa en un documento llamado "Igualdad de oportunidades en el aprendizaje supervisado". Intenta corregir el modelo para que ya no discrimine a grupos específicos de usuarios, basándose en un conjunto de características sensibles.https://fairlearn.github.io/



Modelo de Open data institute

Este es un modelo donde el programador o desarrollador va respondiendo preguntas y al final obtiene un resultado que dice si el algoritmo tiene o tendrá un impacto positivo o no.

AEQUITAS

Fuentes:

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, Rayid Ghani, Aequitas: A Bias and Fairness Audit Toolkit, arXiv preprint arXiv:1811.05577 (2018). (PDF)

Charla Gob Lab UAI: Analizando disparidades en modelos de ML (youtube)

http://www.datasciencepublicpolicy.org/projects/aequitas/

Es una herramienta de código abierto que permite auditar modelos de ML para detectar y mitigar sesgos.

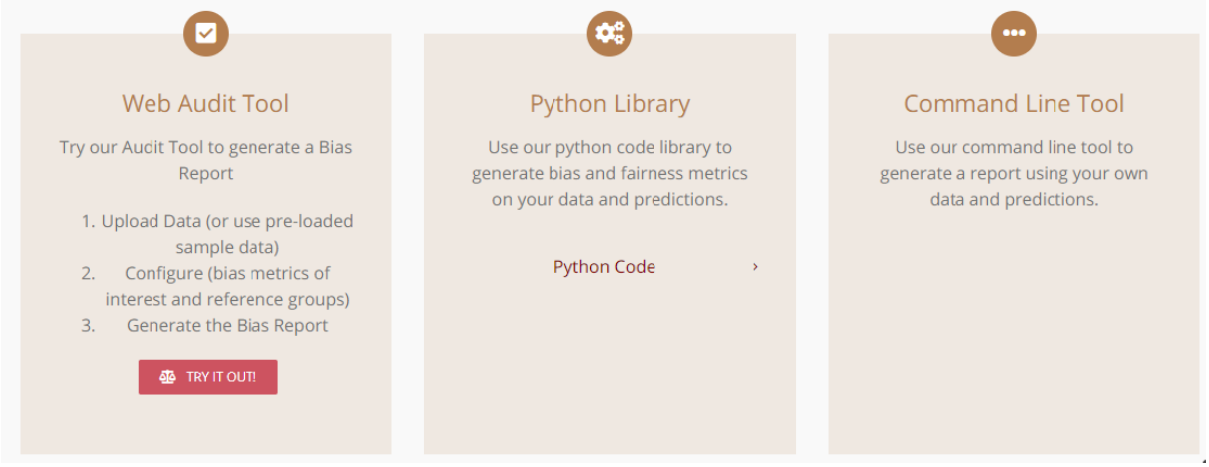
Permite a los usuarios probar modelos para medir distintas métricas de sesgo y definiciones de equidad (fairness) en relación con múltiples grupos y subgrupos de la población, con el objetivo de crear consciencia entre los distintos grupos de interés (stakeholders) respecto a la existencia de bias y de fairness (justicia) como un KPI principal (indicador).
**Un KPI, conocido también como indicador clave o medidor de desempeño o indicador clave de rendimiento, es una medida del nivel del rendimiento de un proceso. El valor del indicador está directamente relacionado con un objetivo fijado previamente y normalmente se expresa en valores porcentuales

La posibilidad de medir, permite optimizar el sistema y habilitar el uso de métodos de mitigación de sesgo o escoger mejor modelos de ML.

La idea es que Aequitas sea una ***herramienta*** de fácil uso para que los científicos de datos y los tomadores de decisiones no tengan excusa para no medir el impacto de los m

Existen 3 maneras de usarlo:

- Desde el sitio web y subir el dataset (no queda guardado)
- Directamente desde una biblioteca de Python
- Command Line tool (línea de comandos)



Desde Triage (*)

https://github.com/dssg/triage

- Definir atributos que consideran importantes (que no se quieren perjudicar o proteger): Age - Sex - Race
- Cuál es el grupo de referencia (favorecido): age: 35-49; sex: male; race: white

Desde Python.

Interfaz orientada a objetos.

3 simples pasos:

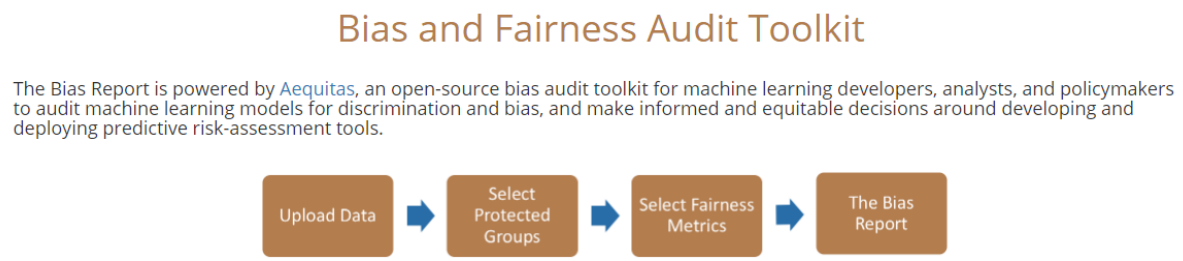
- Define los grupos . Data frame. Se calcula cantidades estadísticas para cada subgrupo
- Calcula las disparidades entre las métricas, toma razones entre las métricas y el bias
- fairness, da reporte sencillo de qué cosas no están siendo justas en el data set

https://github.com/dssg/triage

En forma **extremadamente simplificada**:

- Definir grupo (data frame).
- Se sube el dataset, las predicciones.Se configura la métrica de sesgo para

grupos de atributos de interés protegidos, así como grupos de referencia.
- Se generará un reporte con las métricas.
- Fairness: Luego arrojará en qué métricas hay disparidad según los grupos definidos como protegidos y grupos bases.



Fuente: http://aequitas.dssg.io/

QUÉ SE NECESITA PARA AUDITAR UN MODELO?

PREDICCIONES

Atributos que definen a grupos protegidos. Ej. Etnicidad, sexo, nivel de ingresos

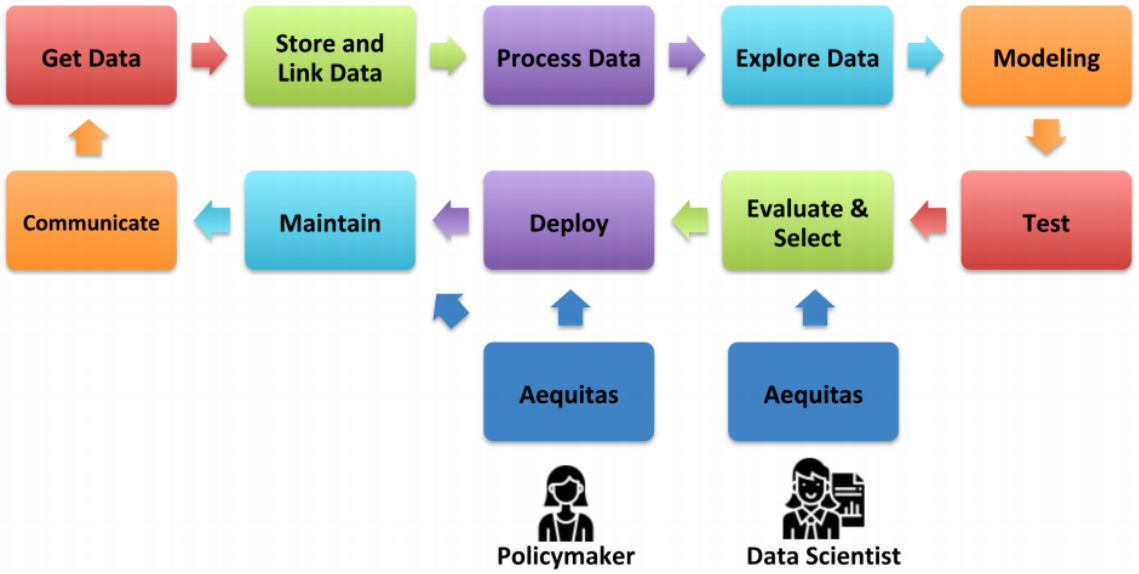
Etiquetas reales. Si hay errores en disparidad. si esta persona estuvo o no en tal escuela, si fallo en ella etc.

El paradigma de la auditoria: *Si no lo puedes medir, no lo puedes mejorar*

Se promueve el uso de Aequitas por dos grupos de actores diferentes, por el científico de datos a la hora de desarrollar modelos de IA y medir el impacto de riesgo (los científicos estarían acostumbrados a seleccionar en base al performance o desempeño del modelo, pero debiera ser en base a qué tan justo es). Y por los tomadores de decisiones, para que antes de aceptar aplicar un modelo de IA o si ya lo utilizan, poder comprender qué sesgos existen en ese modelo y mitigarlos de ser necesario.

Aequitas in the larger context of the ML pipeline. Audits must be carried internally by data scientists before evaluation and model selection. Policymakers (or clients) must audit externally before accepting a model in production as well as perform periodic audits to detect any fairness degradation over time.Audits must be carried internally by data scientists before evaluation and model selection. Policymakers (or clients)

must audit externally before accepting a model in production as well as perform periodic audits to detect any fairness degradation over time.



Audits must be carried internally by data scientists before evaluation and model selection. Policymakers (or clients) must audit externally before accepting a model in production as well as perform periodic audits to detect any fairness degradation over time.
Auditar para bias y fairness es el primer paso para crear conciencia y tomar decisiones más informadas respecto al desarrollo y puesta en producción de modelos de ML, que pueden afectar la vida de los individuos.

Aequitas permite realizar análisis de sesgos y equidad respecto de múltiples atributos y no solo respeto de uno solo pre definido. **Se basa en una definición de sesgo entendida como medida de disparidad entre grupos, en comparación con un grupo de referencia.** A su vez, el grupo de referencia, se puede seleccionar utilizando varios criterios. *Our formulation, allows performing bias and fairness analysis on any multi-valued attribute, and not just for pre-defined protected attributes. We define bias as a disparity measure across groups when compared with a reference group. This reference group can be selected using different criteria.*

Se considera que hay paridad en el impacto, cuando la fracción de elementos del grupo que se predice como positiva, es la misma en todos los grupos. Por otro lado, se considera que hay paridad estadística o demográficamente en los *predictor*, si hay una fracción igual de elementos de cada grupo entre todos los positivos previstos (es decir, se distribuyen todos por igual).

Por lo tanto, definimos el sesgo como una medida de disparidad de los valores métricos de un grupo dado en comparación con un grupo de referencia. Esta referencia se puede seleccionar utilizando diferentes criterios.

Por ejemplo, se podría usar el grupo mayoritario (con mayor tamaño) entre los grupos definidos por A, o el grupo con el mínimo valor de la métrica de grupo, o el enfoque tradicional de fijar un grupo históricamente favorecido (p. ej. raza: blanca).

Las diferentes medidas de equidad varían en importancia para el usuario final según el costo e impacto de la intervención:

- Si las intervenciones son muy caras o pueden perjudicar al individuos, entonces querríamos minimizar los falsos positivos (centrándonos en la tasa de descubrimiento falso y / o Tasa de falsos positivos).

- Si las intervenciones son predominantemente asistenciales, debería preocuparse más por los falsos negativos (centrándose en la tasa de omisiones falsas y / o la tasa de falsos negativos).

- Falsos positivos
- Falsos negativos

Lo distintivo de Aequitas es que tiene presente que bias y fairness no son conceptos absolutos y que están necesariamente vinculados al escenario al que se aplican, como asimismo al análisis y a la interpretación dentro de ese contexto. Por otro lado, pretende ser útil también a autoridades en general y no limitarse a ser entendido únicamente por personas con conocimiento técnico

El criterio de equidad/justicia (fairness) es flexible ya que se basa en un parámetro de valor real para controlar el rango de valores de disparidad que pueden considerarse justos.

Definen dos tipos de injusticia: Supervisada y no supervisada (unsupervised).

El concepto de equidad de Aequitas se basa en un impacto de grupo utilizando restricciones de paridad. Nuestra formulación e implementación
de justicia es flexible, ya que se basa en un parámetro de valor real.

ARQUITECTURA. Se basa en los siguientes componentes:

Input data

Requiere la siguiente data como input:

- Set de predicciones: entidades/grupos? y puntajes dados a esas entidades
- atributos para cada entidad (edad, sexo, etc)
- Resultados/etiquetas para cada entidad

Parámetros de configuración

- Atributos y valores de interés (ej. Género: masculino, femenino). Definir qué atributos se consideran importantes y cuál es el grupo de referencia sobre el cual trabajar
- Valores de referencia para cada grupo, para calcular las proporciones de sesgo (ej. Masculino para género)
- Medidas de sesgo para calcular. La disparidad no será un valor absoluto, dependerá de qué métrica y performance se usará.

Output

Aequitas generará outputs en los siguientes formatos:

- Gráfico de database: Cálculo de los errores en la métrica y la medida de disparidad.
- Reporte en PDF.
- Reporte visual interactivo: Permite al usuario explorar en forma interactiva

Posibles soluciones desde el ámbito Humano

Cualquier tipo de protocolo o regulación es necesario que se considere a los siguientes actores: Sector publico, Sector privado, grandes empresas- pymes- participación ciudadana

Sector publico (Estado y organismos internacionales)

-Gobernanza que cree un marco para la cooperación de las autoridades competentes.
-Marco Juridico que sirva de sustento y protección ante los avances tecnológicos
En el libro blanco del 2020 establece: La estructura de gobernanza debe garantizar la mayor participación de partes interesadas posible. Debe consultarse a las partes interesadas (organizaciones de consumidores e interlocutores sociales, empresas, investigadores y organizaciones de la sociedad civil) sobre la aplicación y futuro desarrollo del marco.

Respecto de las actividades tecnológicas, creemos necesario además que existan códigos de ética y que los mismos sean presentados desde la formación terciaria o universitaria.

Particulares (desarrolladores) Interdisciplinariedad

Junto con la recolección de mejores datos creemos que la interdisciplinariedad es sumamente necesaria. Con filtrado colaborativo, es decir, agregando personas o productos al equipo que no sean afines a nosotros al momento de enseñar a un algoritmo.

ALGUNAS POSIBLES CONCLUSIONES:

- **En un informe publicado la semana pasada (fines de agosto) por el Instituto de Internet de Oxford, los investigadores encontraron que una de las trampas más comunes en las que caen las organizaciones al implementar algoritmos es la creencia de que solucionarán problemas estructurales realmente complejos. Estos proyectos "se prestan a una especie de pensamiento mágico", dice Gina Neff, profesora asociada del instituto y coautora del informe. Los algoritmos no pueden reparar sistemas rotos. Heredan los defectos de los sistemas en los que están ubicados.**

- Los sesgos y las decisiones basadas en sesgos siempre han existido. Los sesgos podrán provenir de los humanos, de los datos (muestra y etiquetado), de los sistemas de inteligencia artificial tanto de las decisiones basadas en los sistemas de ML, como de las acciones posteriores basadas en esas decisiones. (ML: decisiones; acciones)

- Kate Crawford, de Microsoft Research, explicaba en IEEE Spectrum que "es hora de reconocer que los algoritmos son una creación humana que hereda nuestros prejuicios [...]: **nuestra IA será tan buena como lo seamos nosotros".**

- Importancia del factor humano, la ética y los valores defendemos y queremos

- La definición de equidad no dependerá de los científicos de datos, tampoco de las estadísticas, dependerá del contexto en el cual se inserte el sistema. De ahí entonces es que cobra vital importancia el factor humano, la ética y los valores que defendemos y queremos, por un lado, y la necesidad de monitorear constantemente los sistemas de inteligencia artificial por otro. En este sentido, es imperativo que la implementación de sistemas de IA vaya acompañada de auditorías a los modelos, como alguno de los que mencionamos previamente.

- **A pesar de que revisar el concepto de equidad y sus parámetros en el algoritmo antes de implementar un sistema de ML, puede parecer de una necesidad evidente y obvia, aún no es un procedimiento estandarizado. Por eso es importante dar a conocer los desarrollos que**

se están llevando a cabo por aquellos que creen en una IA útil y responsable.

- No existe una definición universalmente aceptada y absoluta de lo que significa que decisiones basadas en un sistema un sistema sean justa

- **Si bien muchos de los modelos de IA hoy utilizado contienen sesgos y han llevado a decisiones discriminatorias, investigaciones han demostrado que las decisiones de los *policy maker* parecieran ser más sesgadas aún y por sobretodo más difíciles de corregir, básicamente por su carácter de seres humanos.** Los sistemas tienden a ser más precisos, con igual o menos sesgos que las personas, pero más "fácilmente" corregibles.

- Los modelos no se insertan en el vacío, son sistémicos e involucran múltiples factores y variantes, por tanto, puede que sean justos,pero el resultado de las acciones que habilita el modelo de ML no lo sea. Importancia de entender el contexto de los sistemas de IA, y del resultado de los mismo.

*** INFORMACIÓN DE INTERÉS SOBRE OTROS SISTEMAS DE AUDITORÍA A SISTEMAS DE ALGORÍTMICOS y otra info de relevancia**

https://www.mckinsey.com/business-functions/mckinsey-analytics/how-we-help-clients

https://algorithmwatch.org/en/jobs/

https://www.technologyreview.es/s/7950/unamonos-para-evitar-la-discriminacion-de-los-algoritmos-que-nos-gobiernan

https://orcaarisk.com/

https://www.technologyreview.es/s/8344/los-algoritmos-sesgados-estan-por-todas-partes-y-parece-que-nadie-le-importa

https://ainowinstitute.org/AI_Now_2019_Report.pdf

https://revista.une.org/11/la-eliminacion-de-los-sesgos-en-los-algoritmos.html