Data Science

www.datascience.pe

Data Science

Sara Iris Garcia

EXPOSITOR

SARA IRIS GARCIA

Data scientist

@montjoile

sarairisgarcia

sarairis.garcia@gmail.com

TEMA: Fairness in Machine Learning

www.datascience.pe

Data Science

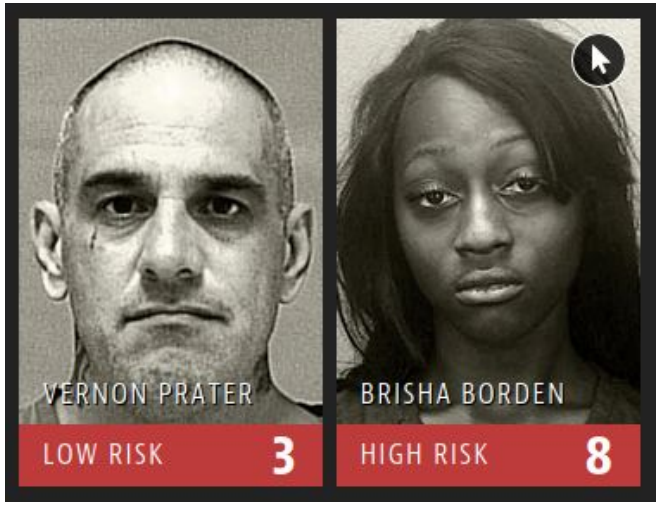
Fairness in Machine Learning

Contenido

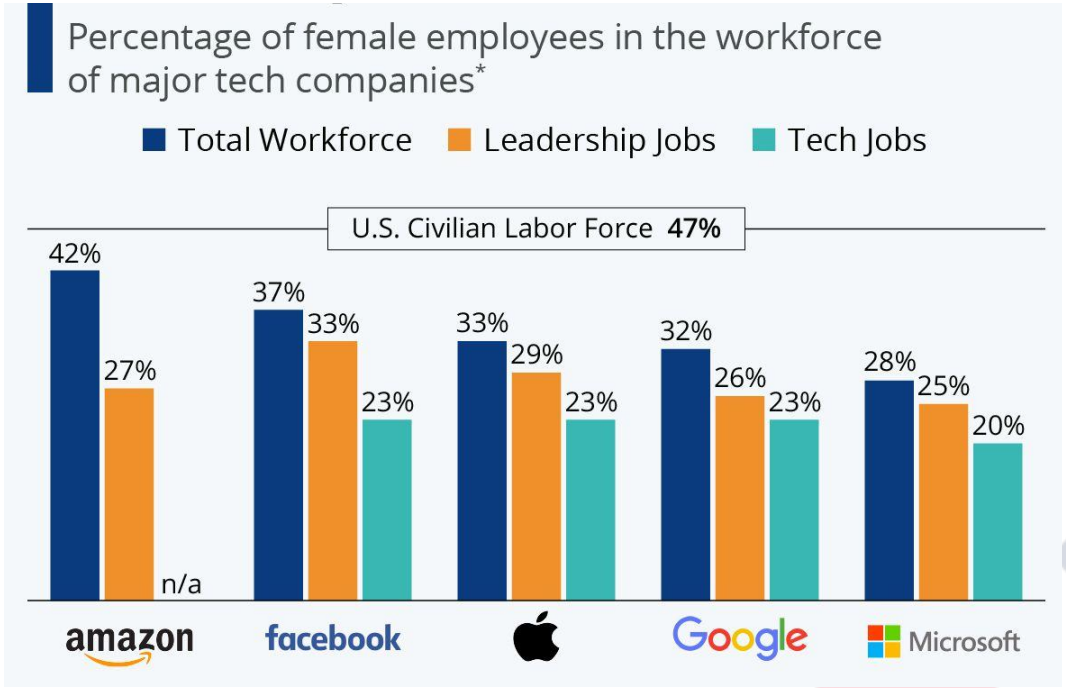
- IA injustas que han tenido impacto real
- Qué es Fairness
- Importancia de la equidad en IA
- Otros conceptos relacionados a la equidad
- De dónde proviene el sesgo
- Tipos de sesgos
- Métricas de Equidad
- Frameworks
- Direcciones de investigación actuales

IA injustas que han tenido impacto real

- IA para contratación de personal descarta a mujeres para puestos técnicos
- IA policiaca implica a afroamericano en un crimen que no cometió



www.propublica.org



www.statista.com

Fuentes:
www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-biasagainst-women-idUSKCN1MK08G
www.seattletimes.com/business/technology/wrongfully-accused-by-an-algorithm
www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

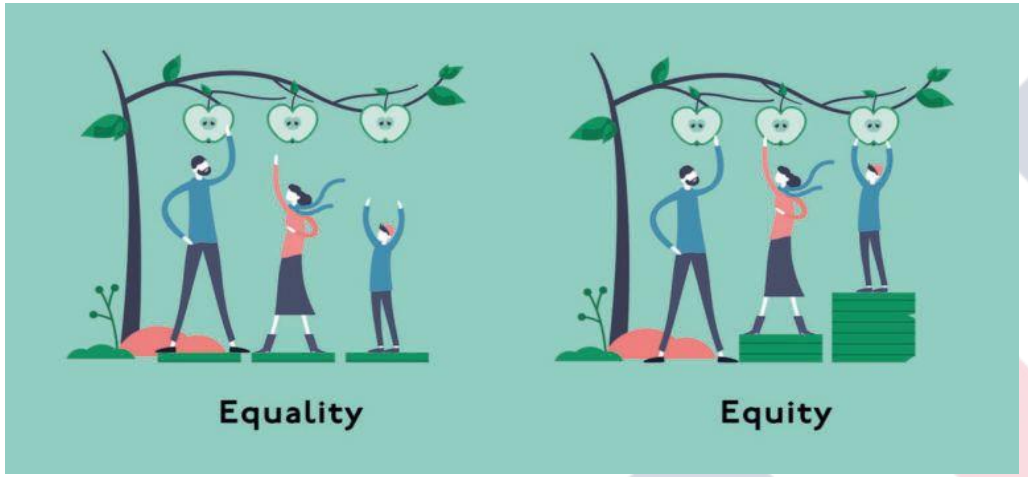
Qué es Fairness o Equidad?

"Los algoritmos y datasets pueden reflejar, reforzar, or reducir prejuicios injustos. Reconocemos que distinguir los prejuicios justos de los injustos no siempre es sencillo y difiere entre culturas y sociedades. Intentaremos evitar impactos injustos en las personas, en particular los relacionados con características tales como raza, etnia, género, nacionalidad, ingresos económicos, orientación sexual, capacidad y creencias políticas o religiosas."

-Principios de IA de Google

"Uno de los mayores desafíos en el desarrollo de algoritmos justos radica en decidir qué significa realmente la justicia."

- Dr Chris Russell, investigador del instituto Alan Turing



Importancia de la Equidad en IA

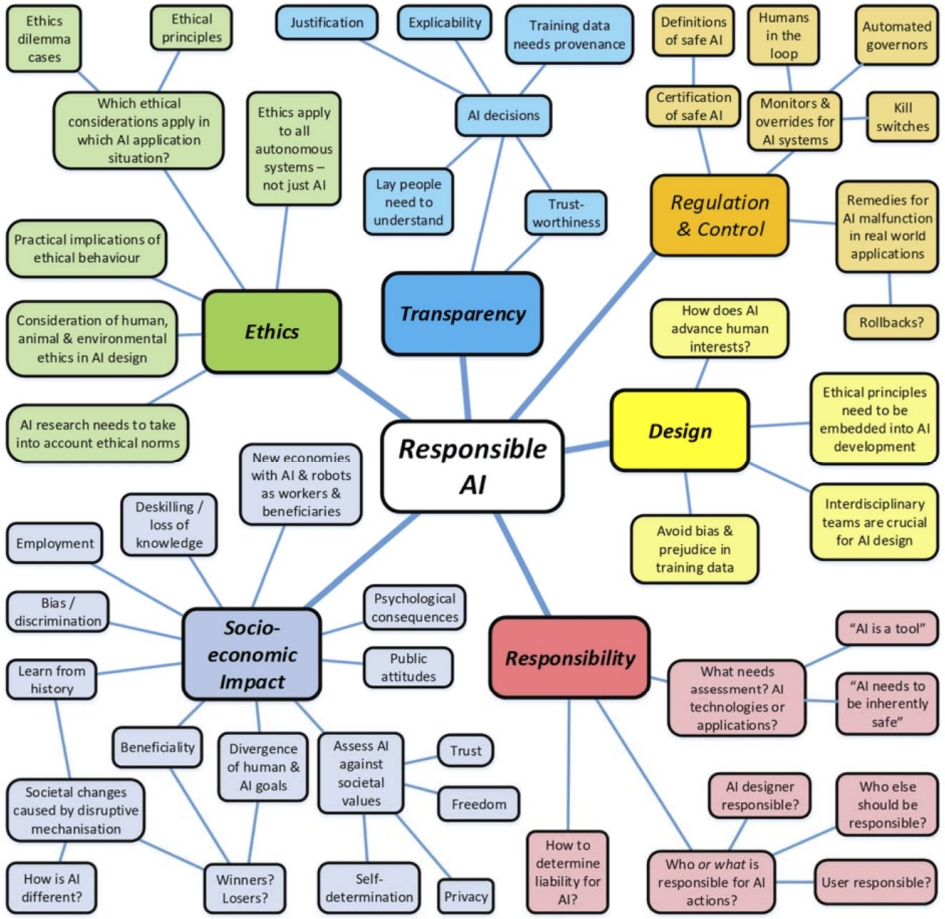
- Regulaciones como GDPR
- Evitar usos malintencionados:
  - Parques en China usan IA para evitar robo de papel higiénico
  - IBM se retira del mercado de reconocimiento facial
  - Amazon y Microsoft no venderán IA al gobierno de USA

www.ipvm.com

www.cnn.com

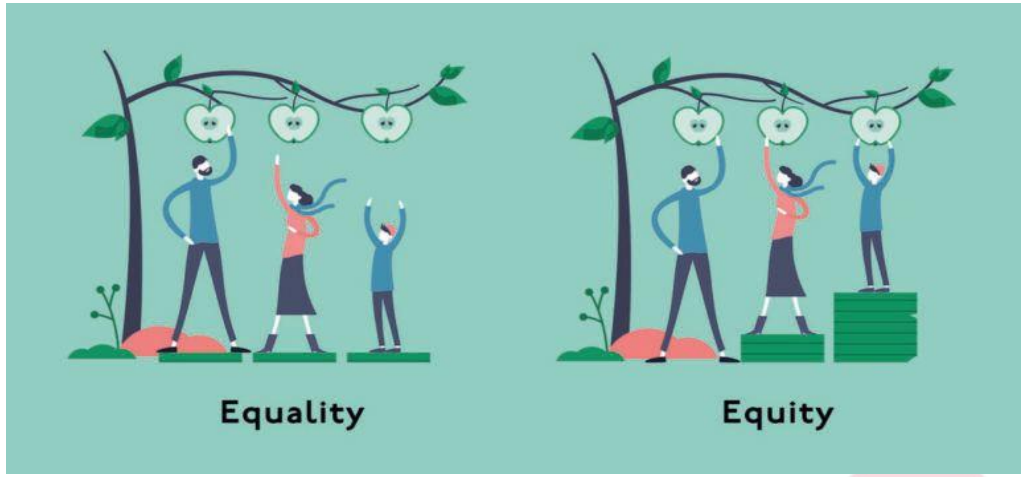
Fuentes:
www.sciencedirect.com/science/article/abs/pii/S026736491830044X
www.bbc.com/news/technology-52978191
www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition
www.washingtonpost.com/technology/2020/06/10/amazon-rekognition-police

Conceptos relacionados a la Equidad



Conceptos relacionados a la Equidad

- IA Responsable
  - Explicable
  - Interpretable
  - Ética
  - Segura
  - Centrada en el ser humano
  - Regulada
  - Justa (equitativa)
  - Fiable
  - Causal
  - Confianza



www.mostly.ia

Conceptos relacionados a la Equidad

IA Responsable

Explicable: que tenga la capacidad de explicar un modelo después de que se ha desarrollado.
Interpretable: que su arquitectura sea transparente y comprensible, que sus parámetros y características de entrada se puedan interpretar.
Transparente: que su explicación sea entendible, que pueda explicar cómo llegó a una determinada decisión.
Ética: que sea capaz de responder a situaciones de forma ética.
Segura: que tenga contramedidas similares contra las amenazas internas y cibernéticas que se verían en el software tradicional.
Centrada en el ser humano: que provea una experiencia efectiva a través de las interacciones del usuario
Regulada: que cumpla con los requisitos reglamentarios relevantes, ya sea con GDPR, CCPA, FCRA, ECOA u otras regulaciones.
Justa: que sea equitativa, no tenga sesgos y no discrimine
Privacidad: que proteja la privacidad de los datos de los que aprende
Fiable: que sea robusta en cualquier escenario de la vida real, que generalice bien, que no sea vulnerable a los ataques adversarios.
Causal: que las asociaciones aprendidas reflejen causas verdaderas en lugar de correlaciones falsas.
Confianza: que sea adecuada por las razones correctas, por ejemplo, que podamos predecir correctamente los límites de decisión del algoritmo.

Data Science

Algorithmic bias - de dónde proviene?

- **Sesgo de la muestra:** ocurre cuando la población está sobrerrepresentada o subrepresentada en un conjunto de datos de entrenamiento.
- **Sesgo de etiqueta:** (o clase) ocurre cuando el proceso de anotación introduce sesgo durante la creación de los datos de entrenamiento.
- **Sesgo de proxy de resultado:** se produce cuando a pesar de que el atributo protegido no sea accesible para el modelo, éste puede aprender a discriminar usando un proxy que esté correlacionado con el atributo protegido.

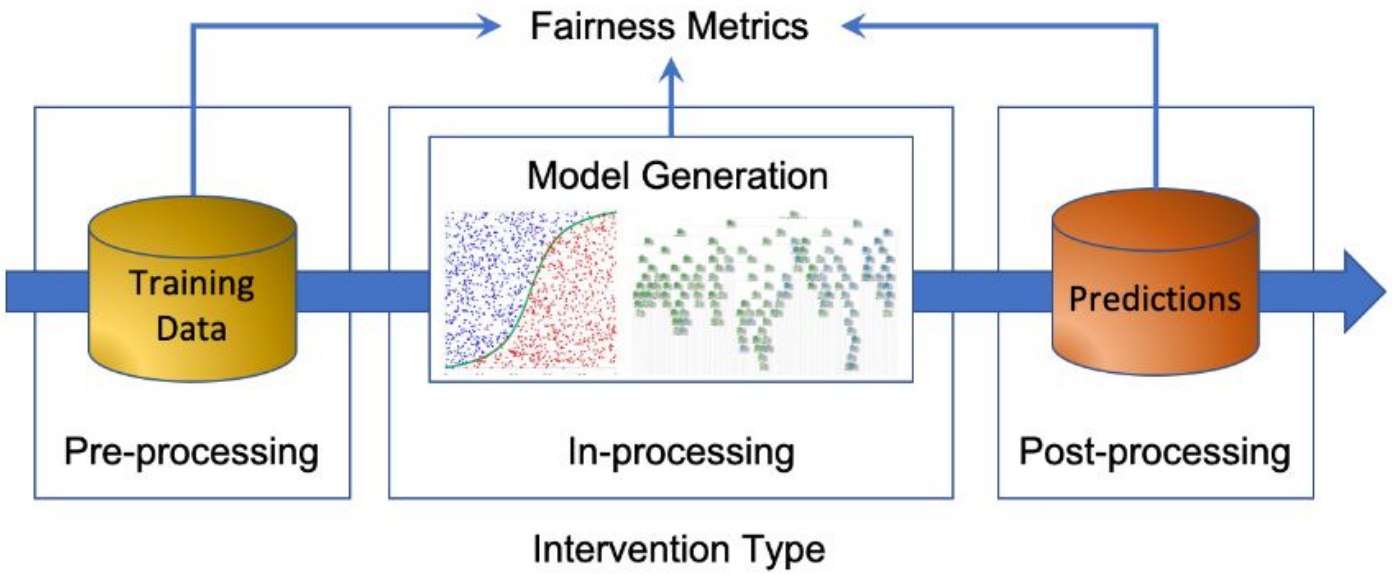
Data Science

Tipos de Sesgos

- Histórico
- Representacional
- De medición
- De evaluación
- De agregación
- De población
- Paradoja de Simpson
- Falacia de datos longitudinales
- De muestreo
- De comportamiento
- De producción de contenido
- De vinculación
- Temporal
- Del observador
- De auto-selección
- Emergente
- Social
- De presentación
- Algorítmico
- De causa-efecto
- De variable omitida
- De ranqueo
- De interacción del usuario
- De financiamiento

Métricas de Equidad

Equidad en el proceso de construcción de IA:



Caton, Simon & Haas, Christian. (2020). Fairness in Machine Learning: A Survey.

Métricas de Equidad

- **De acuerdo al individuo o instancia:**
  - **Métricas de Grupo**
    - Métricas basadas en paridad
    - Métricas basadas en la matriz de confusión
  - **Equidad Individual**
    - Fairness through awareness
- **Razonamiento Causal**
  - Causal inference for social discrimination reasoning
  - Causal Reasoning for Algorithmic Fairness
  - Avoiding Discrimination through Causal Reasoning
- **Equidad Bayesiana**
  - Bayesian Models and Algorithms for Fairness and Transparency
  - Fair Bayesian Optimization
  - Bayesian Fairness
  - Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference

www.yields.io

Métricas de Equidad

- **Equidad Contrafáctica**
  - Counterfactual Reasoning for Fair Clinical Risk Prediction
  - Counterfactual Fairness in Text Classification through Robustness
  - Counterfactual Fairness
- **Equidad Representativa**
  - Learning Fair Representations
  - Fairness in representation: quantifying stereotyping as a representational harm
  - Costs and Benefits of Fair Representation Learning

McNamara, Ong and Williamson (AIES 2019)

Gráficos interactivos:

- Igualdad de oportunidades vs paridad demográfica
- Causal Bayesian Networks

Data Science

Mitigación de Sesgos

- **Aprendizaje adversario**
  - Adversarial Learning for Counterfactual Fairness
  - Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction
  - Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations
  - Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations
- **Calibración**
  - On Fairness and Calibration
  - Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints
  - Equality of Opportunity in Supervised Learning
- **Incorporación de priors**
  - Incorporating Priors with Feature Attribution on Text Classification
  - Mitigating Gender Bias in Captioning Systems
- **Recopilación de datos**
  - Why Is My Classifier Discriminatory?
  - Incorporating Dialectal Variability for Socially Equitable Language Identification
  - REPAIR: Removing Representation Bias by Dataset Resampling
- **Mitigación de Representación**
  - Learning Gender-Neutral Word Embeddings
  - Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Librerías de código abierto

FAIRML: framework para auditar de modelos ML predictivos de caja negra https://pypi.org/project/fairml/0.1.1.5rc2/	`pip install fairml` https://towardsdatascience.com/evaluating-machine-learning-models-fairness-and-bias-4ec82512f7c3
Aequitas: toolkit para auditar sesgos y equidad https://github.com/dssg/aequitas	`pip install aequitas` http://aequitas.dssg.io/
Themis-ml: librería para ML que implementa varias métricas de equidad compatible con sklearn. https://themis-ml.readthedocs.io/en/latest/	`pip install themis-ml` https://facctconference.org/static/tutorials/bantilan_themis18.html
What-If Tool (WIT): plugin para Jupyter Notebook que provee una interface para visualizar el comportamiento de modelos ML. Accesible como feature en TensorBoard y parte de Fairness-indicators https://github.com/PAIR-code/what-if-tool	`pip install wit-widget` https://pair-code.github.io/what-if-tool/get-started/
FairLearn: librería de Python para evaluar la equidad de modelos ML y mitigar sesgos https://fairlearn.github.io/	`pip install fairlearn` https://github.com/fairlearn/fairlearn/tree/master/notebooks
Responsibly: toolkit para auditar y mitigar el sesgo y la equidad para modelos ML compatible con sklearn https://github.com/ResponsiblyAI/responsibly	`pip install responsibly` https://docs.responsibly.ai/demos.html
AI Fairness 360 (AIF360): toolkit para verificar sesgos en datasets y modelos ML https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/	`pip install aif360` https://github.com/Trusted-AI/AIF360/tree/master/examples

Librerías de código abierto

Arena: dashboard para exploración de modelos ML https://arena.drwhy.ai/docs/	`pip install -U dalex` https://medium.com/responsibleml/python-has-now-the-new-way-of-exploring-xai-explanations-4248846426cf
EthicML: libreria para realizar y evaluar equidad algoritmica https://github.com/predictive-analytics-lab/EthicML	`pip install ethicml` https://wearepal.ai/EthicML/
EthicalML/XAI: toolbox para explicabilidad para ML https://github.com/EthicalML/xai	`pip install xai` https://ethicalml.github.io/xai/index.html
Audit-AI: librería de Python para detectar diferencias demográficas en modelos ML. Compatible con sklearn. https://github.com/pymetrics/audit-ai	`pip install audit-AI` https://github.com/pymetrics/audit-ai/tree/master/examples

Lista completa:
https://techairesearch.com/most-essential-python-fairness-libraries-every-data-scientist-should-know/

Data Science

Frameworks

- LIME
- GRADCAM
- LRP
- DeepLIFT
- Network dissection
- DeepExplain
- FairSight
- Algofairness/Fairness-comparison
- ML-fairness-gym

Direcciones de investigación actuales

- No existe una definición única y universal de equidad que sea aplicable para cualquier contexto.
- Algunas definiciones de equidad son matemáticamente imposibles de satisfacer simultáneamente e incluso presentan conflictos entre sí.
- La literatura se enfoca mayormente a aprendizaje supervisado, con énfasis en problemas de clasificación binaria. Se necesitan enfoques mucho más diversos.
- Faltan herramientas para identificación de variables protegidas o sensibles
- Se necesitan métricas con enfoque enfocado a la industria.

Referencias

- https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/
- https://krvarshney.github.io/pubs/MahoneyVH2020.pdf
- https://scihub.wikicn.top/10.1145/3194770.3194776
- https://arxiv.org/pdf/1910.10045.pdf
- https://arxiv.org/pdf/1906.05684.pdf
- http://cs-people.bu.edu/sameki/ResponsibleAI.html
- https://arxiv.org/pdf/1910.10045.pdf
- https://arxiv.org/pdf/2010.04053.pdf
- https://arxiv.org/pdf/1908.09635.pdf
- igualdad de oportunidad https://arxiv.org/pdf/1610.02413.pdf
- https://developers.google.com/machine-learning/glossary/fairness#fairness_metric
- individual fairness http://dagstuhl.sunsite.rwth-aachen.de/volltexte/2020/12018/pdf/LIPIcs-FORC-2020-2.pdf
- individual fairness https://arxiv.org/pdf/2006.11439.pdf
- individual fairness https://arxiv.org/abs/1104.3913
- https://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdfhttps://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf
- Causal fairness papers list: https://github.com/debmandal/causality-fairness
- Glosario de métricas de equidad: https://developers.google.com/machine-learning/glossary/fairness#fairness_metric
- https://fairmlbook.org/
- Lista de recursos: https://github.com/datamllab/awesome-fairness-in-ai#mitigation-of-machine-learning-models
- Problemas con métricas de equidad:
- https://5harad.com/papers/fair-ml.pdf
- https://arxiv.org/pdf/1609.07236.pdf
- https://link.springer.com/article/10.1007/s12599-020-00650-3

Data Science

Preguntas?

 @montjoile

 sarairis.garcia@gmail.com

Slides:
github.com/montjoile/neurips2020-peru

Data Science

www.datascience.pe