FACULTAD DE
INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

2023

SMART + SUSTAINABLE

Liderar la construcción
de un mundo sostenible

Ética en Ciencia de Datos e IA: Conceptos para un desarrollo responsable

Workshop on Machine and Statistical Learning with Applications

FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

SMART + SUSTAINABLE

Liderar la construcción **de un mundo sostenible**

Ética en Ciencia de Datos e IA: Conceptos para un desarrollo responsable

Reinel Tabares Soto

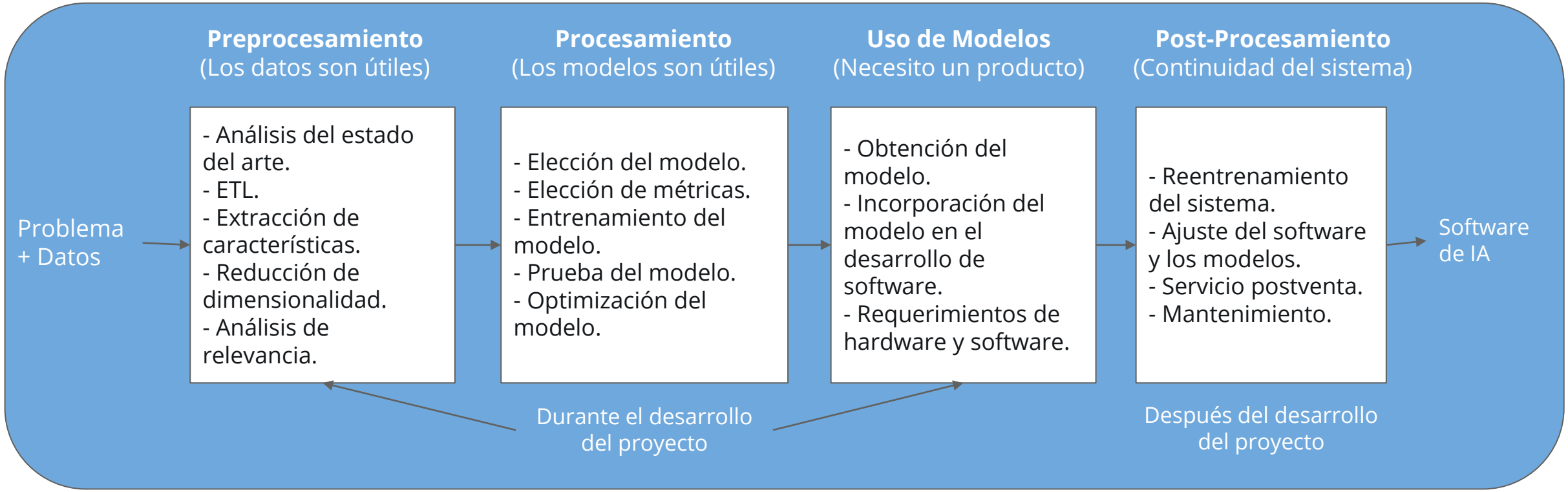
Investigador Facultad de Ingeniería y Ciencias - Ph.D.

Junio 2023

Agenda

1. Introducción
2. Análisis de Impacto
3. Análisis de sesgo y equidad con Aequitas
4. Transparentar modelos con Model Cards
5. Análisis de contrafactuales con What-If Tool
6. Algunas aplicaciones en Chile (IPS, DPP, FONASA)
7. Conclusiones y recomendaciones

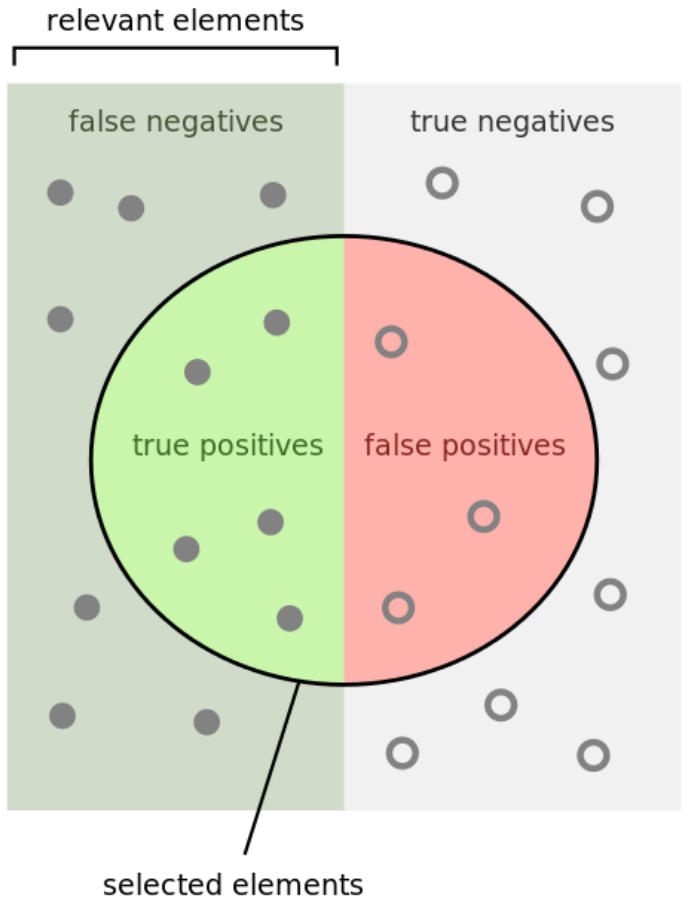
Diagrama de flujo de un sistema de IA



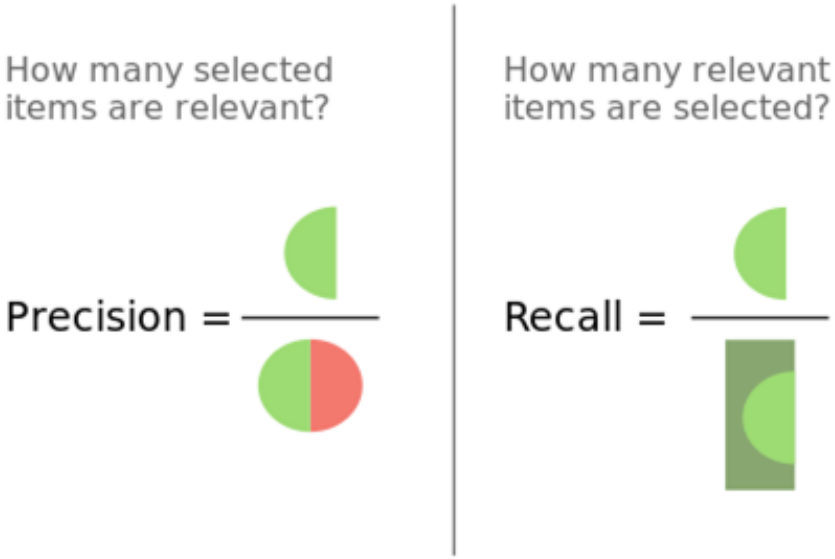
Métricas de rendimiento un modelo de IA

Métricas de rendimiento un modelo de IA

Predicción / Realidad	Tiene Plomo	No tiene Plomo
Tiene Plomo	Verdadero Positivo (VP o TP)	Falso Negativo (FN)
No tiene Plomo	Falso Positivo (FP)	Verdadero Negativo (VN o TN)



Métricas de rendimiento un modelo de IA



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy: Es cuantas casas con plomo y sin plomo predije correctamente del total.

Precisión: Es cuantas casas que yo dije que tenían plomo que realmente tienen plomo con respecto a las que yo dije que tenían plomo.

Recall: Es cuantas casas con plomo acerté de todas las casas con plomo que había.

Ejemplo Práctico

Predicción / Realidad	Tiene Plomo	No tiene Plomo
Tiene Plomo	500 (TP)	1000 (FN)
No tiene Plomo	500 (FP)	10000 (TN)

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Predicción / Realidad	Tiene COVID	No tiene COVID
Tiene COVID	2000 (TP)	10 (FN)
No tiene COVID	500 (FP)	2000 (TN)

Accuracy = 10500 ÷ 12000 = 87.5%
Precision = 500 ÷ 1000 = 50%
Recall = 500 ÷ 1500 = 33.3%

Accuracy = 4000 ÷ 4510 = 88.7%
Precision = 2000 ÷ 2500 = 80%
Recall = 2000 ÷ 2010 = 99.5%

Desbalance vs Sobreentrenamiento

¿Tiene plomo? (predicción)	¿Tiene plomo? (real)
NO	NO
NO	NO
NO	NO
NO	NO
NO	NO
NO	NO
SI	NO
NO	NO
NO	NO
NO	SI

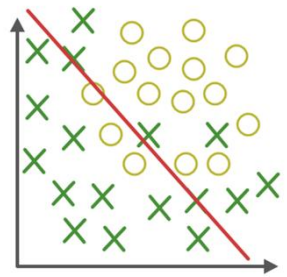
Predicción / Realidad	Tiene plomo	No tiene plomo
Tiene plomo	0 (TP)	1 (FN)
No tiene plomo	1 (FP)	8 (TN)

$$Precision = 0 \div 1 = 0\%$$
$$Recall = 0 \div 1 = 0\%$$
$$Accuracy = 8 \div 10 = 80\%$$

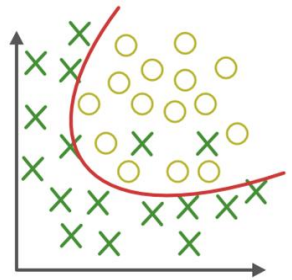
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

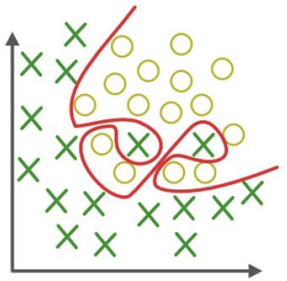
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



Under-fitting
(too simple to explain the variance)

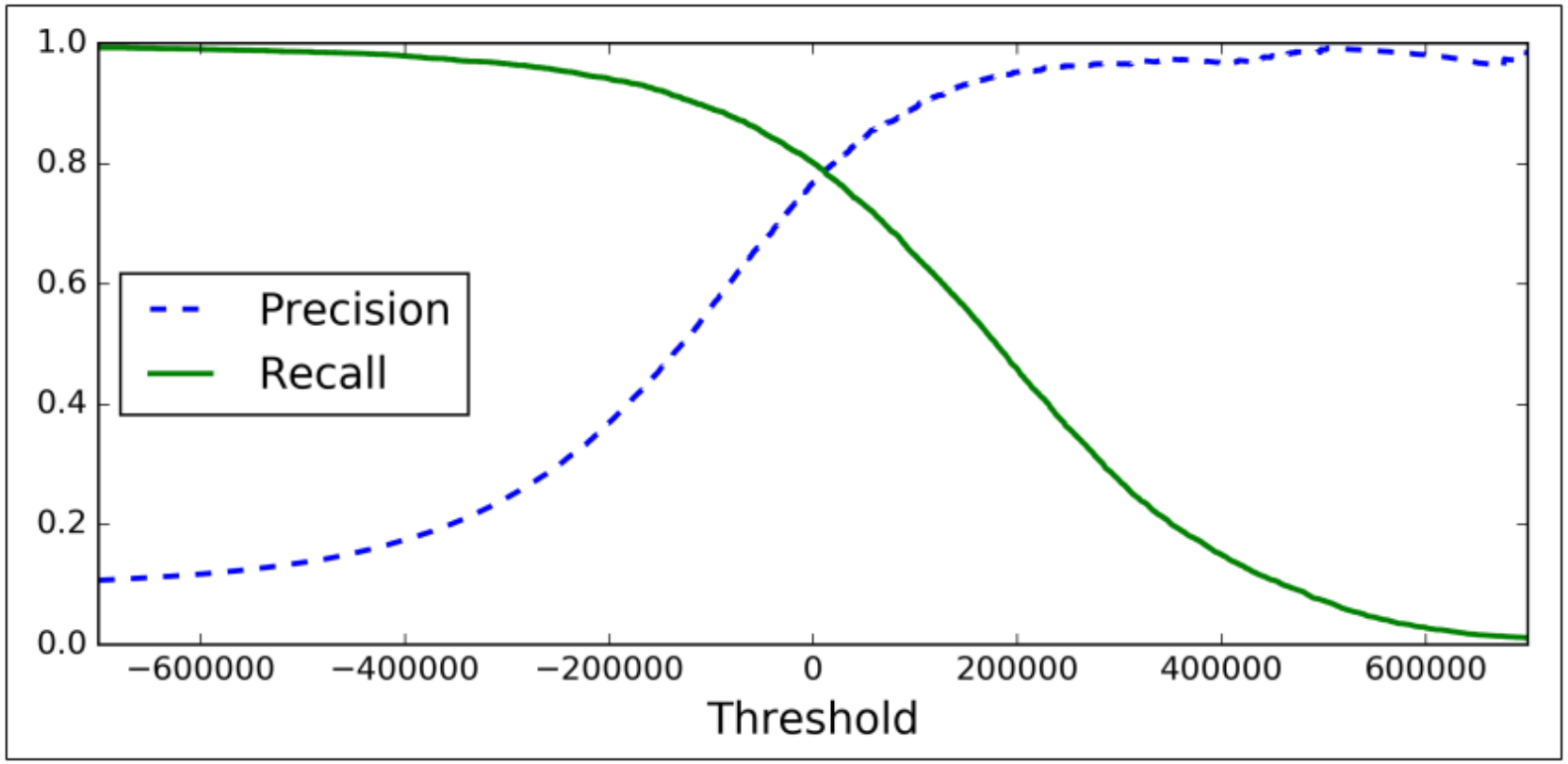


Appropriate-fitting



Over-fitting
(forcefitting--too good to be true)

Precision y Recall



Medida de desempeño de clasificadores

F - measure

- También conocido como F-score.
- Media armónica de precisión y recall.
- Precisión y recall son igual de importantes.

$$F_{measure} = 2 \frac{precision \cdot recall}{precision + recall}$$

Aspectos éticos en IA

Pilares éticos y los principios de la ética de datos

Revisión de 84 marcos de referencia (Jobin et Al, 2019)

Principio	Palabras asociadas
Transparencia	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justicia y Equidad	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution

From Table 2 – Ethical principles identified in existing AI guidelines (Jobin et Al, 2019)

Principio	Palabras asociadas
No maleficiencia	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsabilidad	Responsibility, accountability, liability, acting with integrity
Privacidad	Privacy, personal or private information
Beneficencia	Benefits, beneficence, well-being, peace, social good, common good

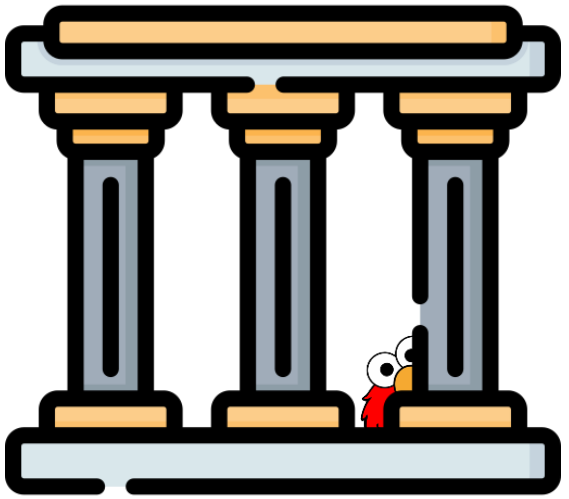
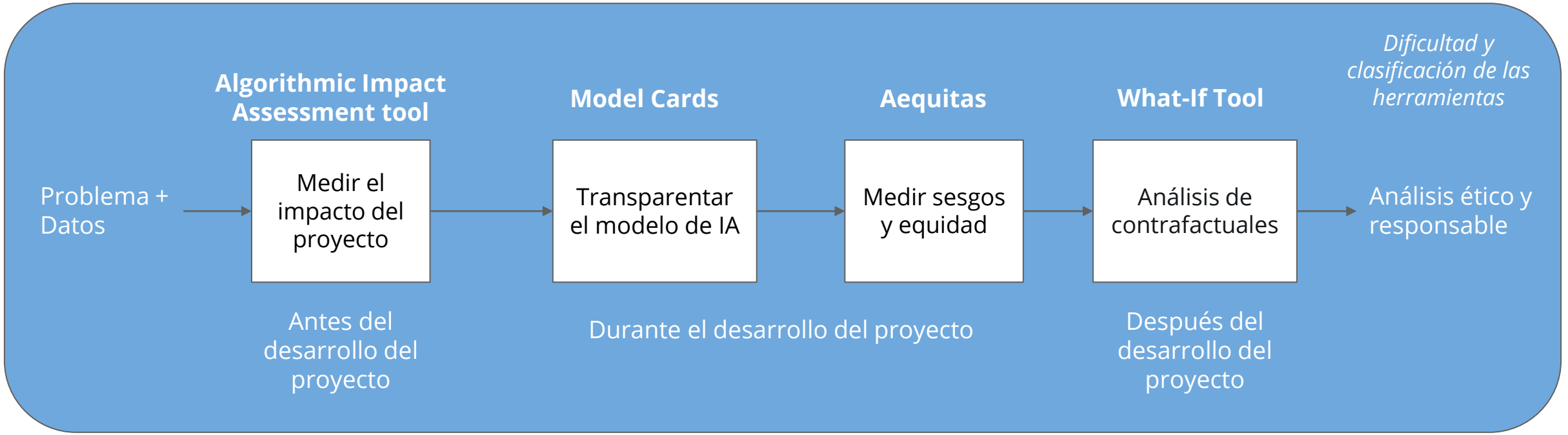


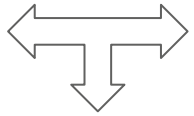
Diagrama de flujo de análisis ético en un sistema de IA



Protección de datos

Protección de datos

RGPD



Normativa Europea

LOPD
Ley Orgánica de Protección de Datos

¿LOPD qué es y cómo cumplirla? | VIU (universidadviu.com)

Ley-19628 28-AGO-1999 MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA - Ley Chile - Biblioteca del Congreso Nacional (bcn.cl)

Normativa Chilena

Análisis de Impacto

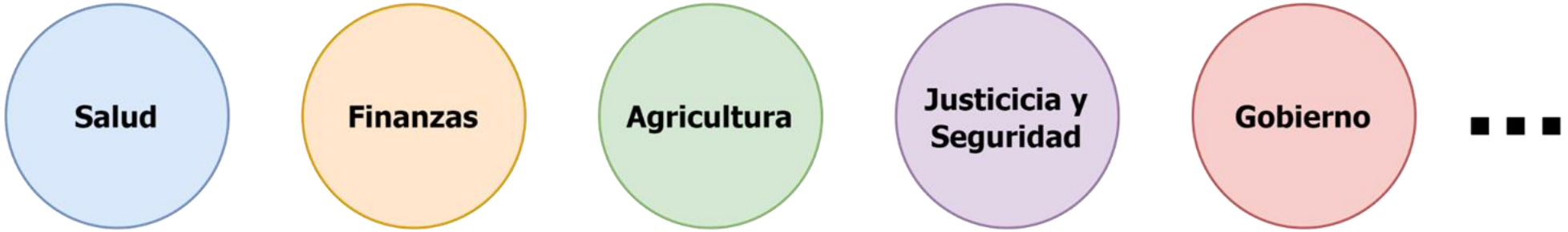
¿Qué es un análisis de impacto?

Proceso que evalúa los posibles efectos positivos y negativos de aplicar técnicas de ciencia de datos o IA a un problema o contexto específico. El análisis de impacto puede ayudar a identificar los beneficios y los riesgos de usar sistemas basados en IA, así como las medidas para mitigarlos.

¿Por qué es importante evaluar el impacto?

¿Por qué es importante evaluar el impacto?

La ciencia de datos y la IA ya no son campos aislados, convivimos a diario con ellos.



IA en diferentes sectores/campos

Salud

Finanzas

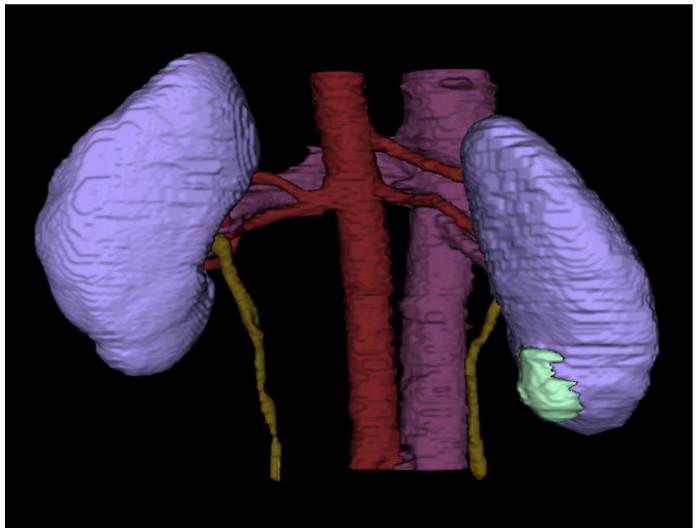
Agricultura

Justicia y Seguridad

Gobierno

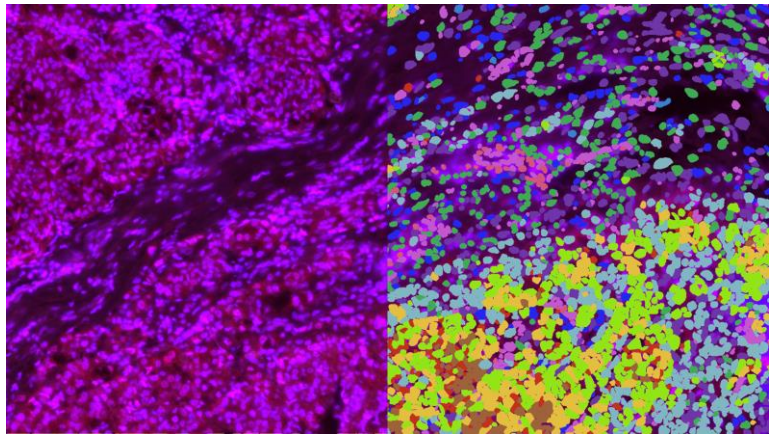
...

Segmentación de tumores



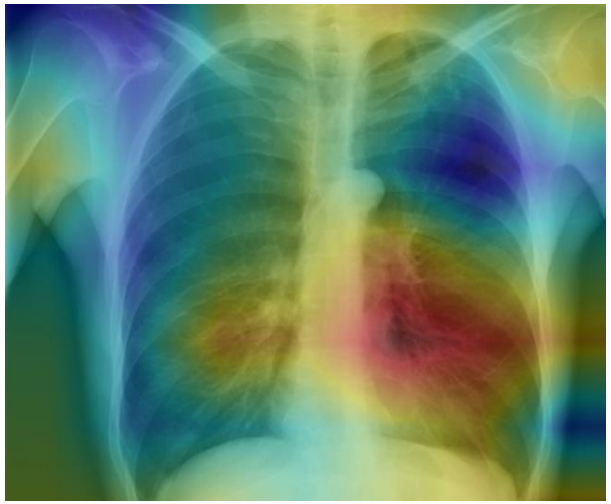
KiTS21

Segmentación de células

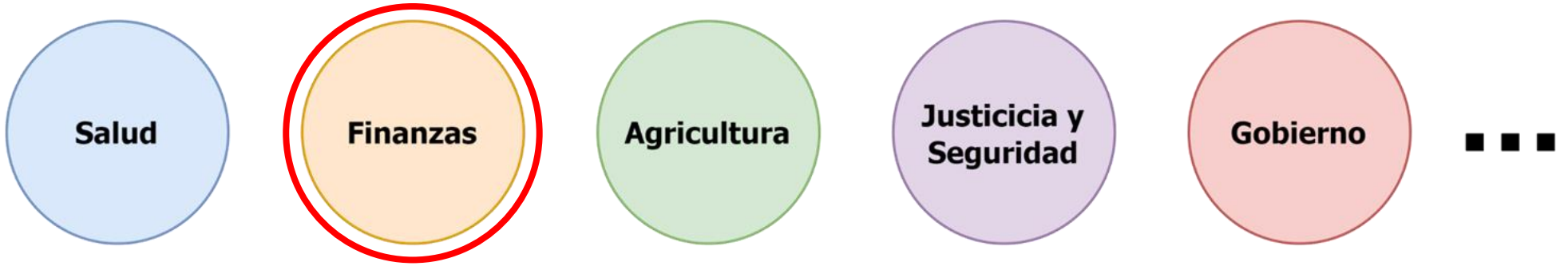


MERSCOPE FFPE

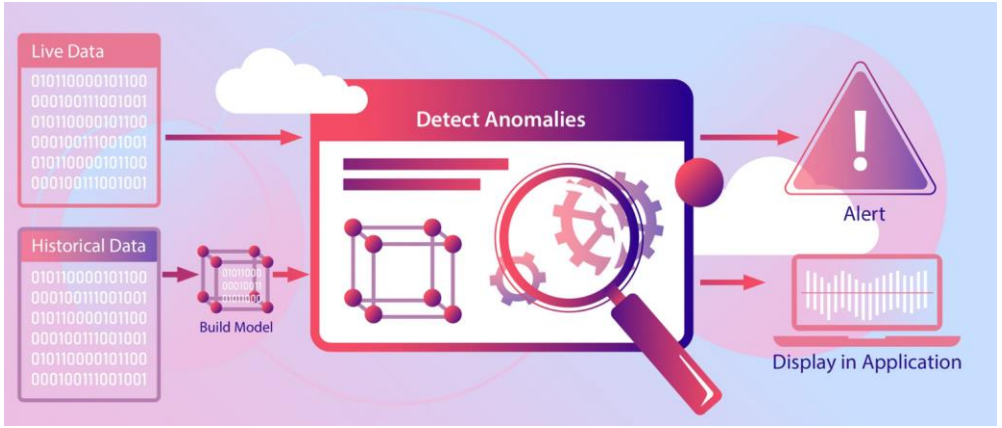
Detección de Neumonía



CheXNet



Detección de fraude



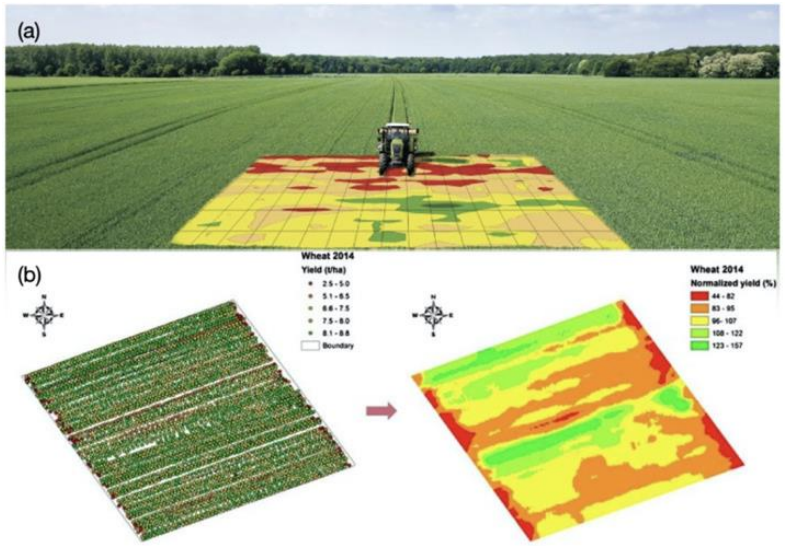
Deep learning for detecting fraud

Predicción del mercado de valores

Artificial intelligence in finance

Salud | Finanzas | **Agricultura** | Justicia y Seguridad | Gobierno | ...

Mapeo de rendimiento



Implementation of artificial intelligence in agriculture

Detección de maleza

Weed Detection – Tensorfield Agriculture

Salud

Finanzas

Agricultura

Justicicia y Seguridad

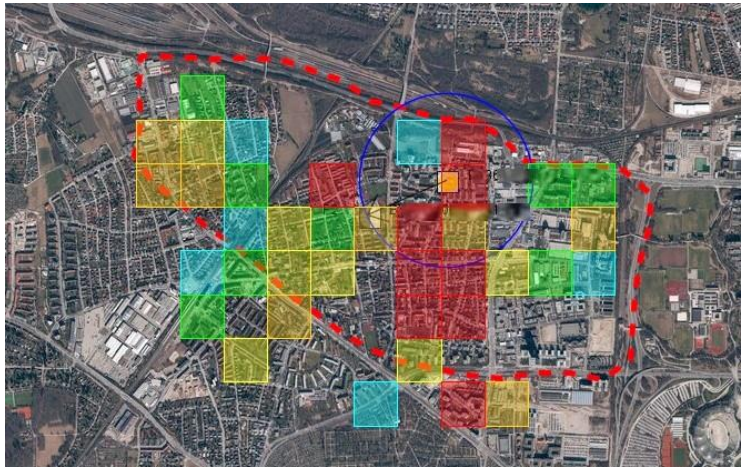
Gobierno

...

Reconocimiento Facial

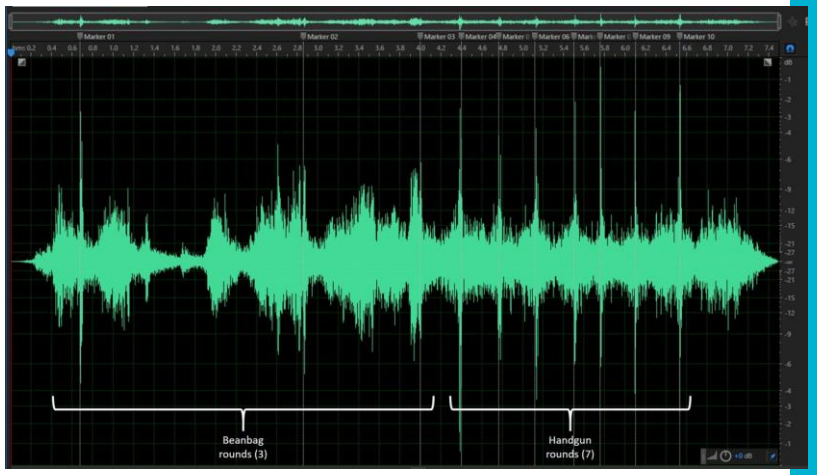
[One Month, 500,000 Face Scans - China]

Predicción de Crímenes

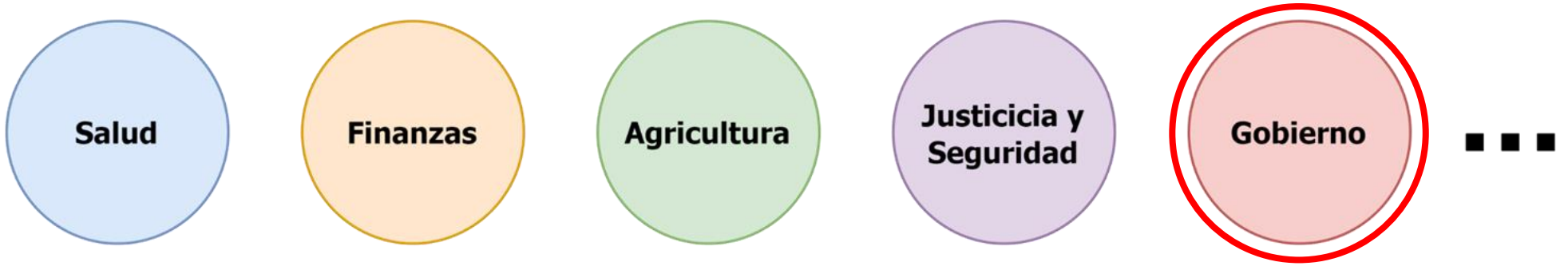


[PRECOBS]

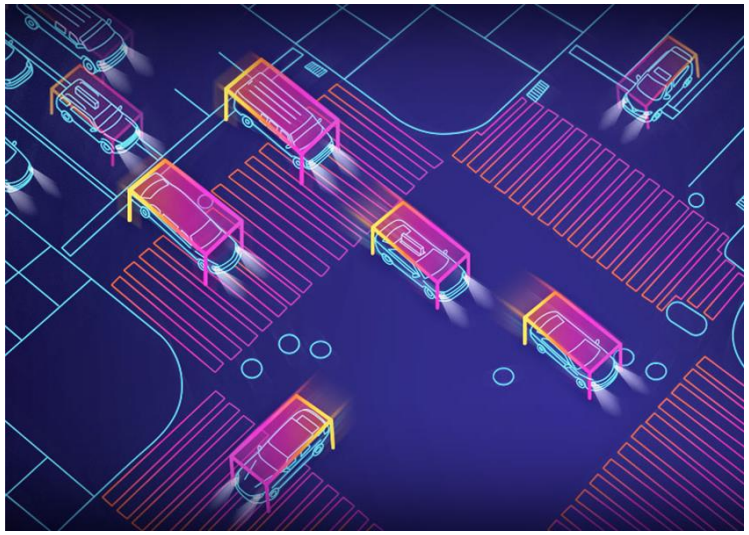
Detección de Armas



[Audio Analysis of Gunshots]

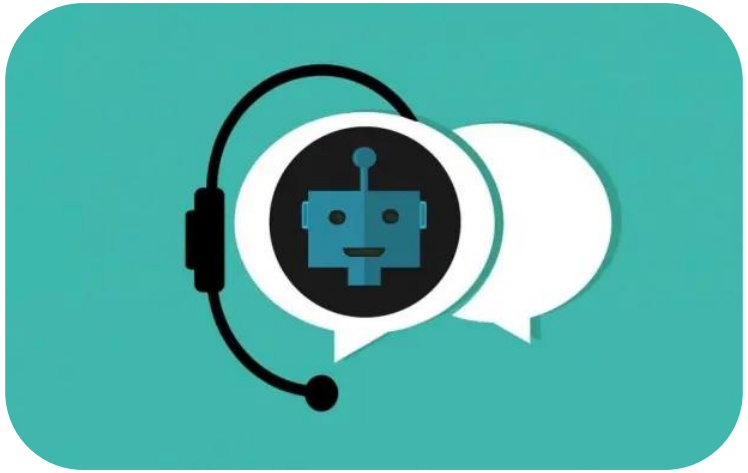


Detección de Tráfico



[FLIR Intelligent Transportation Solutions]

Chatbots



[Chatbot Mona]

Asignación de ayudas

[Machine learning en servicios sociales]

Es necesario anticipar el impacto social a largo plazo y los usos inesperados de la tecnología que creamos hoy. Lo último que deseamos es ser sorprendidos por un futuro trágico que ayudamos a crear [1].

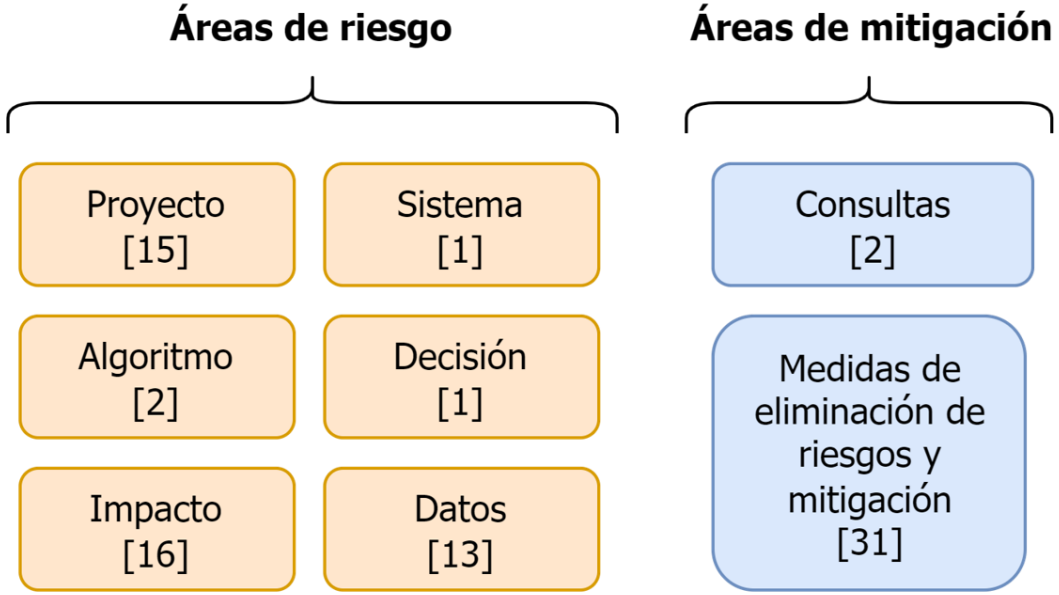
Algorithmic Impact Assessment Tool

Government of Canada

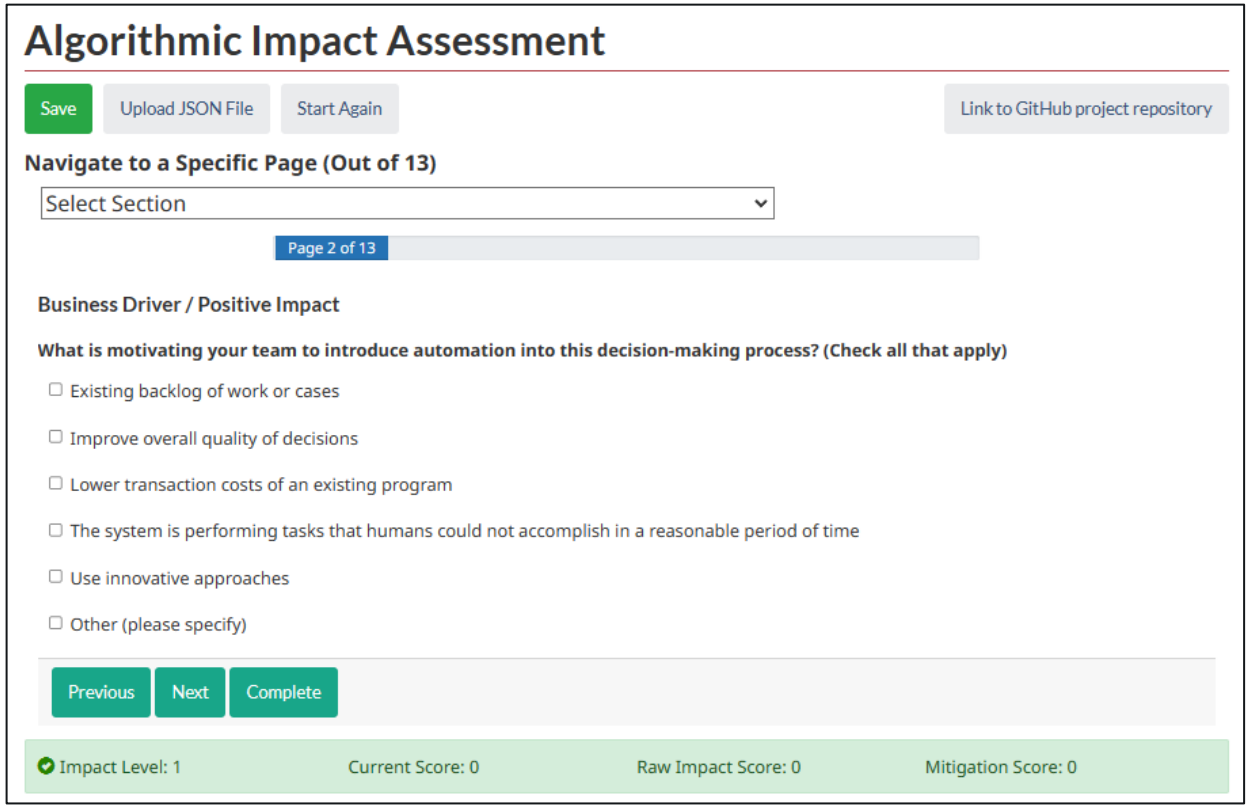
[Algorithmic Impact Assessment Tool - Canada.ca]

Cuestionario que determina el nivel de impacto de un sistema de decisión automatizado.

Los puntajes de evaluación se basan en factores como el diseño del sistema, el algoritmo, el tipo de decisión, el impacto y los datos.



Algorithmic Impact Assessment Tool



Section 1: Impact Level : 2

Current Score : 45

Raw Impact Score: 45

Risk Area	No. of Questions	Project Score	Maximum Score
Risk Profile	4	10	13
Project Authority	1	0	2
About the Algorithm	2	0	6
About the Decision	1	1	6
Impact Assessment	10	14	36
About the Data - A. Data Source	11	20	38
About the Data - B. Type of Data	2	0	6
RAW IMPACT SCORE	31	45	107

Mitigation Score: 32

Mitigation Area	No. of Questions	Project Score	Maximum Score
Consultations	4	2	2
De-Risking and Mitigation Measures - Data Quality	10	10	14
De-Risking and Mitigation Measures - Procedural Fairness	17	19	25
De-Risking and Mitigation Measures - Privacy	4	1	4
MITIGATION SCORE	35	32	45

Identifica riesgos y evalúa los impactos en una amplia gama de áreas:

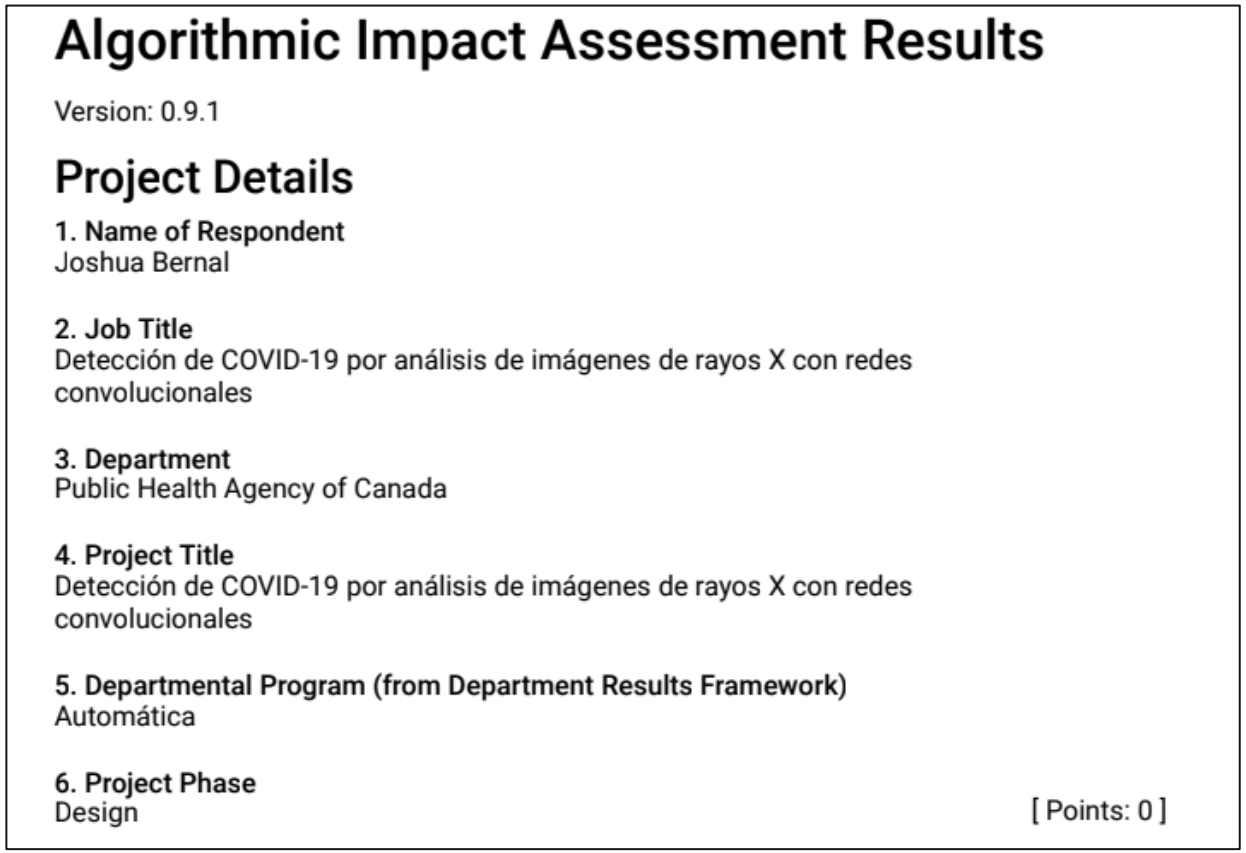
- Los derechos de los individuos o las comunidades.
- La salud o el bienestar de individuos o comunidades.
- Los intereses económicos de individuos, entidades o comunidades.
- La sostenibilidad continua de un ecosistemas

Nivel de Impacto	**Definición**	**Rango porcentual de puntuación**
Nivel I	Poco o ningún impacto	0% a 25%
Nivel II	Impacto moderado	26% a 50%
Nivel III	Alto impacto	51% a 75%
Nivel IV	Muy alto impacto	76% a 100%

Los niveles de impacto se distinguen en función de criterios de reversibilidad y duración esperada.

Ejemplo - Algorithmic Impact Assessment Tool

Enlace a ejemplo de Evaluación de Impacto para modelo ML de detección de Covid-19



Algorithmic Impact Assessment Results

Version: 0.9.1

Project Details

1. Name of Respondent
Joshua Bernal

2. Job Title
Detección de COVID-19 por análisis de imágenes de rayos X con redes convolucionales

3. Department
Public Health Agency of Canada

4. Project Title
Detección de COVID-19 por análisis de imágenes de rayos X con redes convolucionales

5. Departmental Program (from Department Results Framework)
Automática

6. Project Phase
Design [Points: 0]

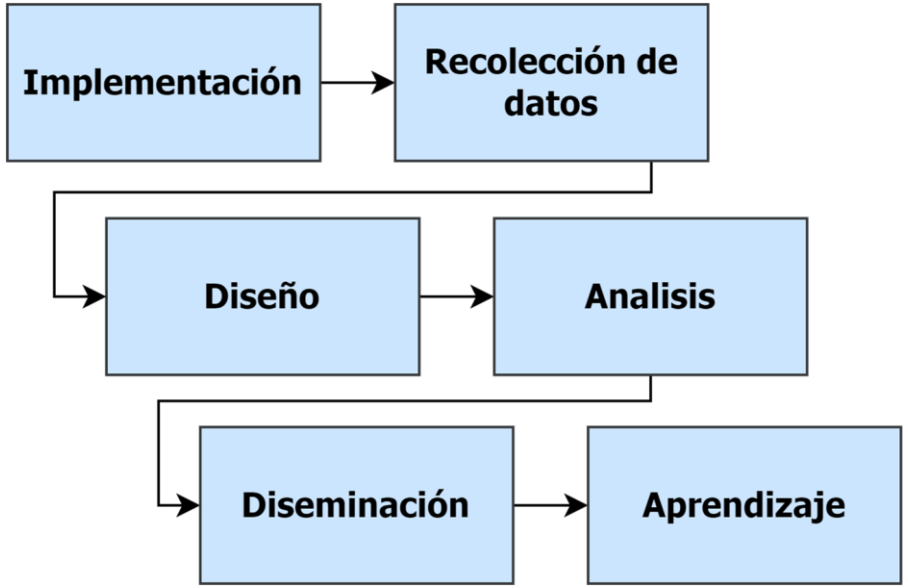
Portal de Evaluación de Impacto

BID Banco Interamericano de Desarrollo

Portal Evaluación de impacto - BID

Herramientas para implementar evaluaciones de impacto. El contenido está organizado según el ciclo de actividades necesarias para completar una evaluación satisfactoria.

Los recursos no solo aplican para proyectos de Ciencia de Datos o Inteligencia Artificial.



Portal de Evaluación de Impacto

Bienvenido al portal de evaluación de impacto

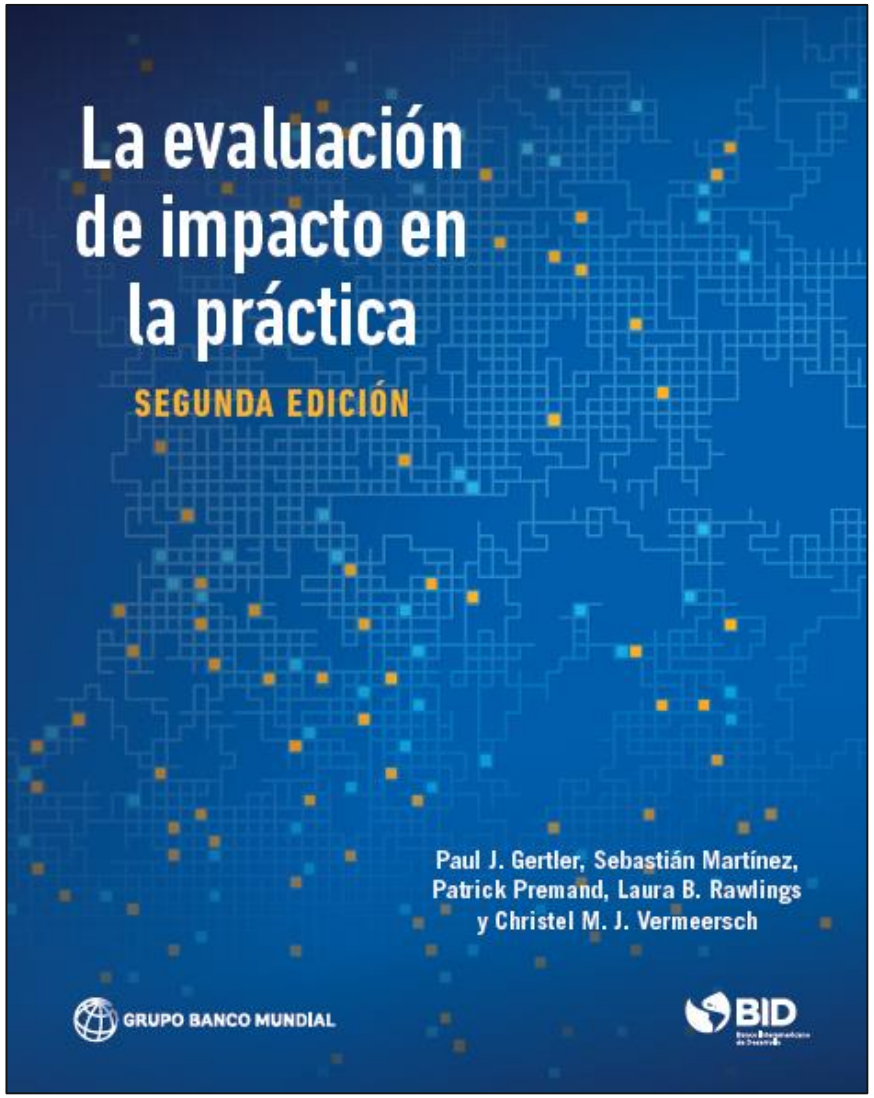
Este portal contiene herramientas para implementar evaluaciones de impacto. El contenido se organiza según el ciclo de actividades necesarias para completar una evaluación satisfactoria.

Diseño	Implementación	Recolección de datos	Análisis	Diseminación	Aprendizaje

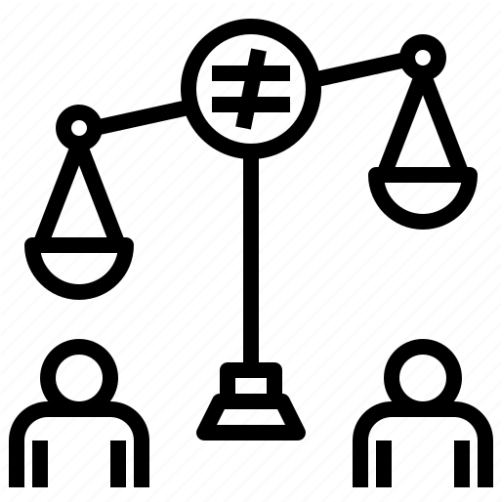
Datos de buena calidad son un insumo clave para las evaluaciones de impacto. Los materiales e instrucciones que se presentan a continuación pueden adaptarse a sus necesidades para el levantamiento de información. También puede consultar nuestro manual del diseñador de cuestionario y el de entrada de datos para obtener más información sobre la preparación de su encuesta.

LEVANTAMIENTO DE DATOS

HOGARES COMUNIDADES UNIDADES DE SALUD ESCUELAS AGRIC



Análisis de Sesgo y Equidad

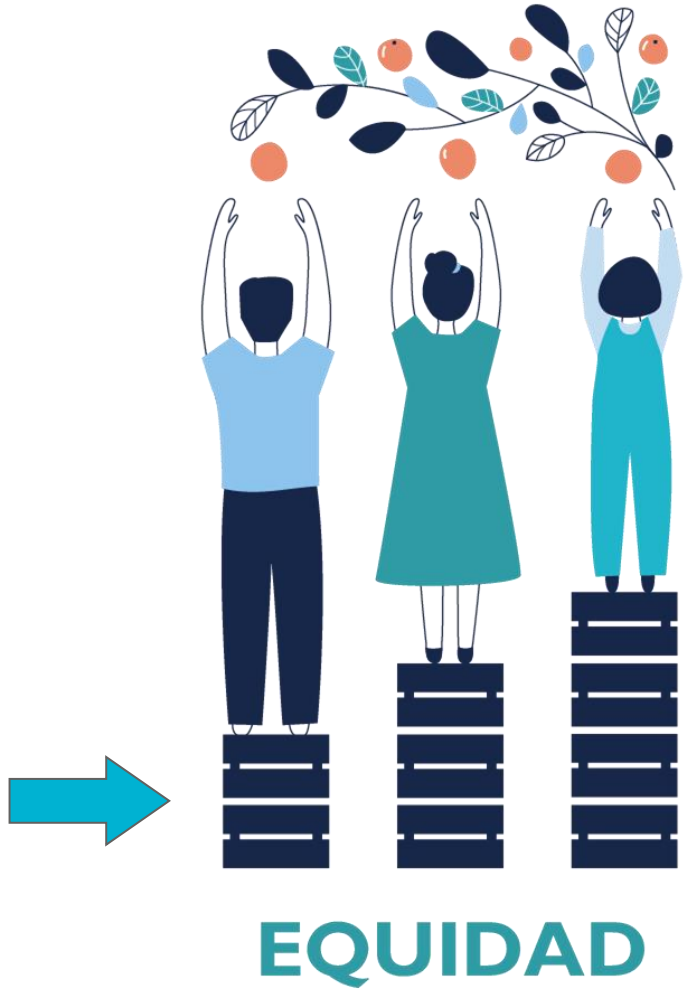


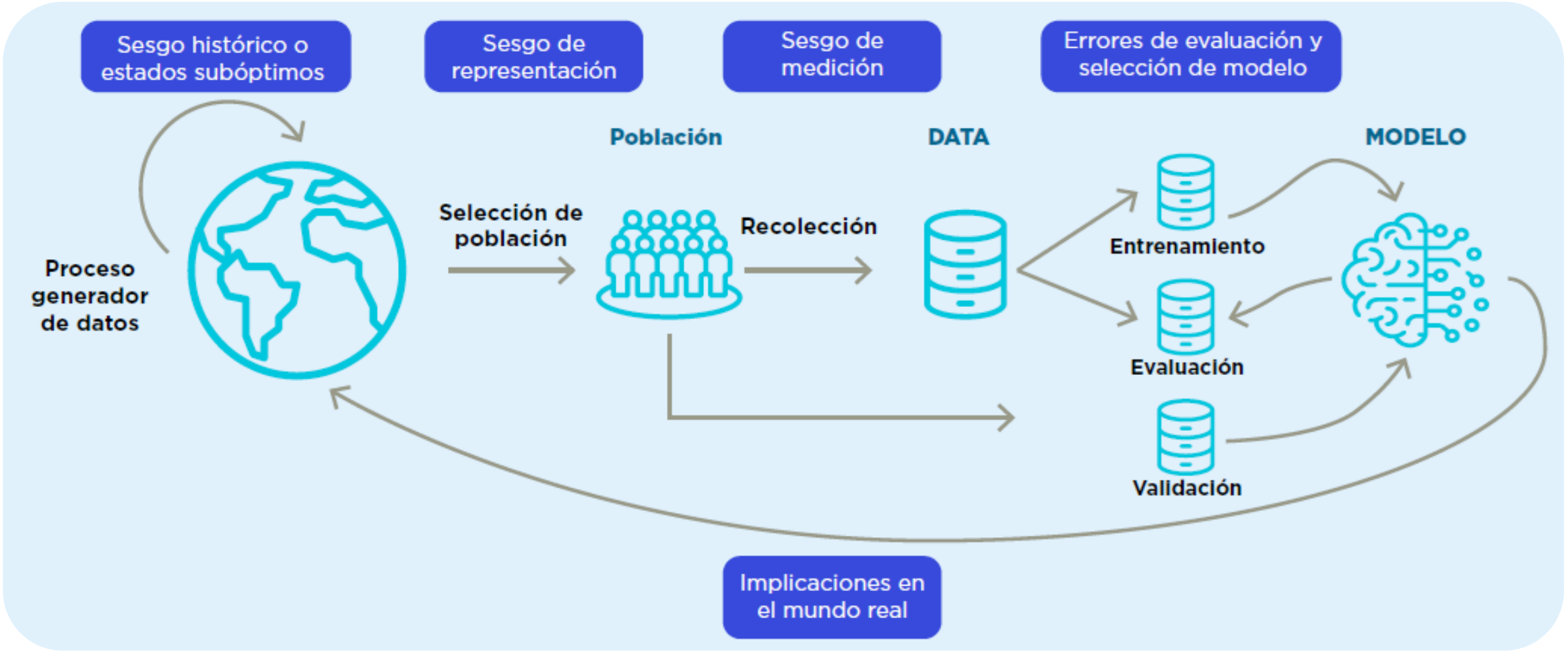
Sesgo y Equidad

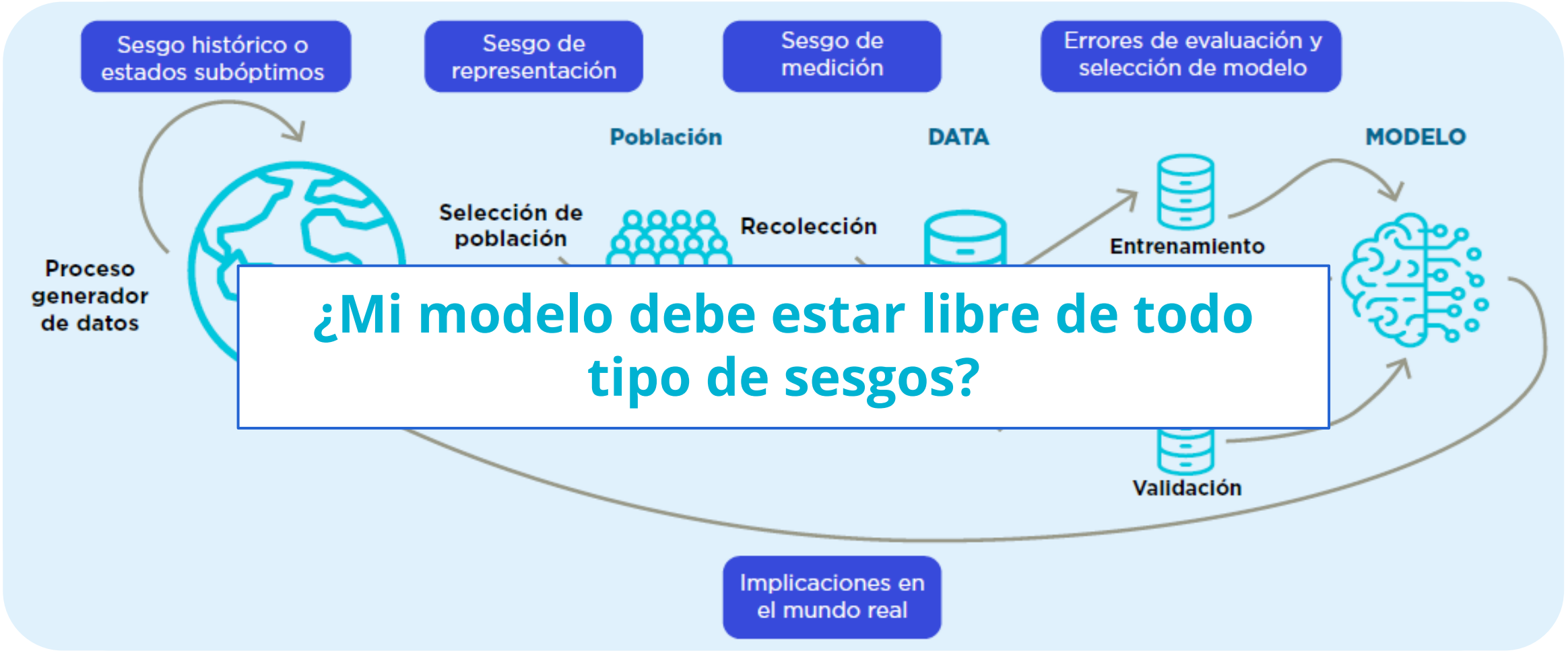
El error del sistema es la diferencia entre el valor predicho, y el valor real de la variable que se está estimando. Si el error es sistemático en una dirección o en un subconjunto específico de los datos, se llama **sesgo** [2].

Cualidad de tratar a todas las personas de manera justa e imparcial, sin discriminación ni favoritismo, y asegurando que tengan acceso a las mismas oportunidades y recursos.

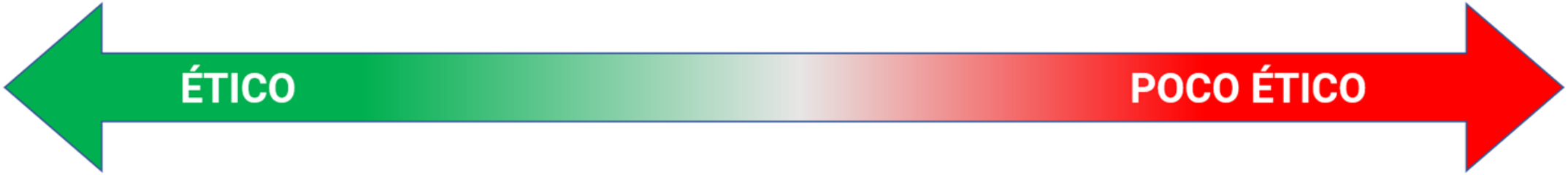
Resultados sesgados.




Sesgo histórico o estados subóptimos
Sesgo de representación
Sesgo de medición
Errores de evaluación y selección de modelo
Población
DATA
MODELO
Proceso generador de datos
Selección de población
Recolección
Entrenamiento
Evaluación
Validación
Implicaciones en el mundo real



¿Mi modelo debe estar libre de todo tipo de sesgos?



Tipo de Sesgo	Intencional (Positivo)	No Intencional (Evitar, aunque sea positivo)	No Intencional (Evitar, no es positivo)	Intencional (Muy grave, debe prevenirse)
Impacto en la Equidad	Los 'sesgos' proporcionan resultados ideales y son necesarios para aplicar los requisitos comerciales de manera ética. Estos son buenos y están destinados a que la empresa brinde resultados justos y justificables a los clientes.	'Sesgos' creados por desconocimiento sin responsabilidad directa o conocimiento por parte de la empresa. Proporcionan beneficios justos sin ningún daño; pero no puede ser explicado o justificado.	Sesgos creados sin conocimiento ni responsabilidad directa por parte de la empresa, pero que resultan en decisiones injustas o discriminación de los clientes. Puede pasar desapercibido y causar daño continuamente. Sin control de la empresa.	Sesgos creados deliberadamente para el 'juego sucio' poco ético para favorecer un resultado que es una discriminación poco ética, injustificable e injusta de los clientes. El peor tipo de sesgo, puede ser ilegal.

Bias and Fairness - Medium

Definiciones de Equidad en ML

- **Equidad Individual**
  - Equidad a través del Conocimiento
  - Equidad a través del Desconocimiento
  - Equidad Contrafactual
- **Equidad de Grupo**
  - Paridad Demográfica
  - Paridad Estadística condicional
  - Igualdad de Probabilidades
  - Igualdad de Oportunidades
  - Igualdad de Trato
  - Equidad de Prueba
- **Equidad de Subgrupo**
  - Equidad en Dominios relacionales

*Definiciones más utilizadas para la **equidad** en los problemas de clasificación algorítmica.*

A Survey on Bias and Fairness in Machine Learning | ACM

¿Qué es el sesgo en IA?

género	raza	promedio	examen admisión	admitido?
hombre	caucásico	0.35	0.91	1
mujer	afroamericano	0.89	0.93	0
hombre	afroamericano	0.84	0.65	0

Caractersiticas X | etiqueta Y

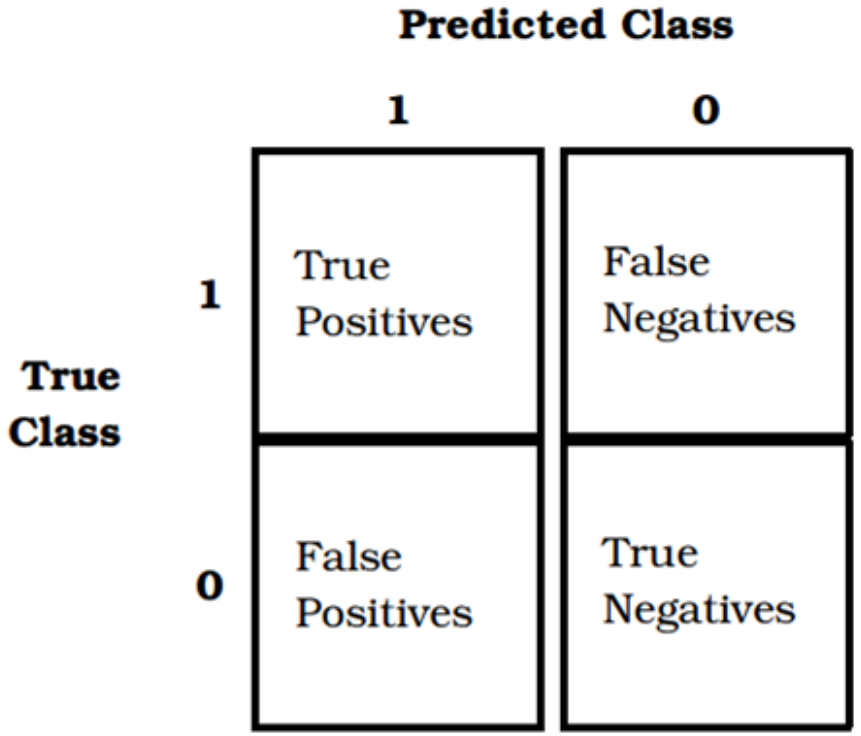
¿Qué es el sesgo en IA?

género	raza	promedio	examen admisión	admitido?
hombre	caucásico	0.35	0.91	1
mujer	afroamericano	0.89	0.93	1
hombre	afroamericano	0.84	0.65	1

atributos protegidos A | Caractersiticas X | etiqueta Y

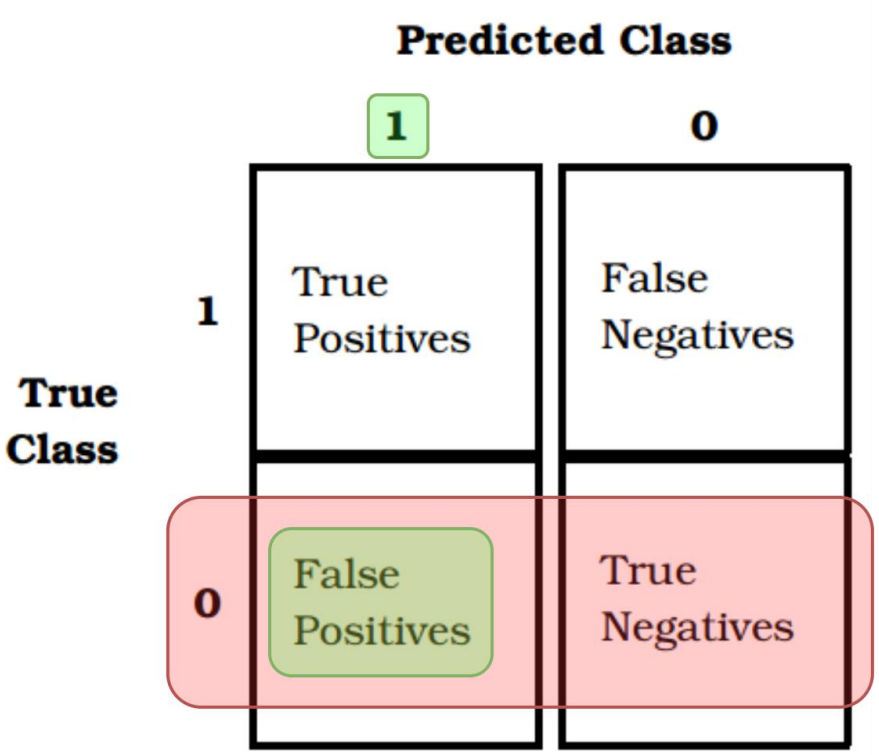
Métricas de Sesgo y Equidad

Matriz de Confusión



	Model Predicted	True Real
True Positive (TP)	1	1
False Positive (FP)	1	0
True Negative (TN)	0	0
False Negative (FN)	0	1

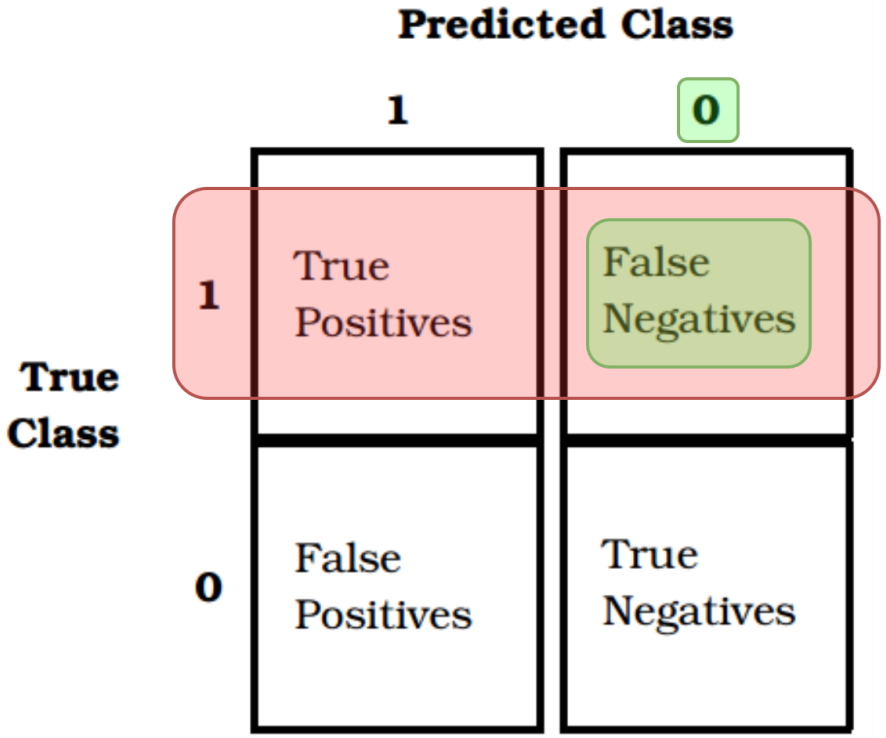
Tasa de Falsos Positivos - False Positive Rate (FPR)



$$FPR = \frac{FP}{FP + TN}$$

Fracción de individuos con etiquetas reales negativas que el modelo clasifica erróneamente con una etiqueta predicha positiva.

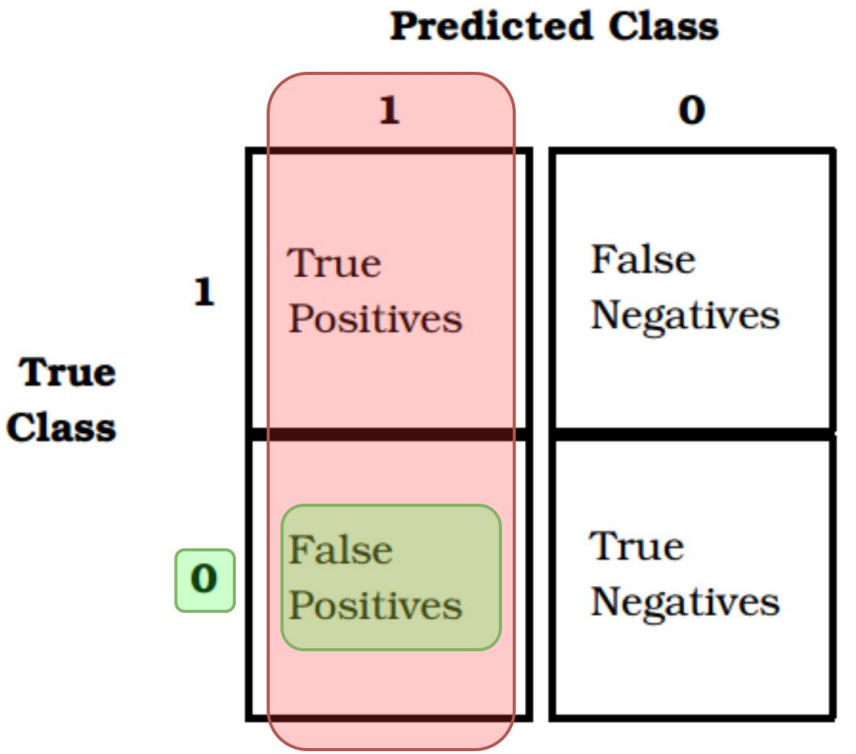
Tasa de Falsos Negativos - False Negative Rate (FNR)



$$FNR = \frac{FN}{FN + TP}$$

Fracción de individuos con etiquetas reales positivas que el modelo clasifica erróneamente con una etiqueta predicha negativa.

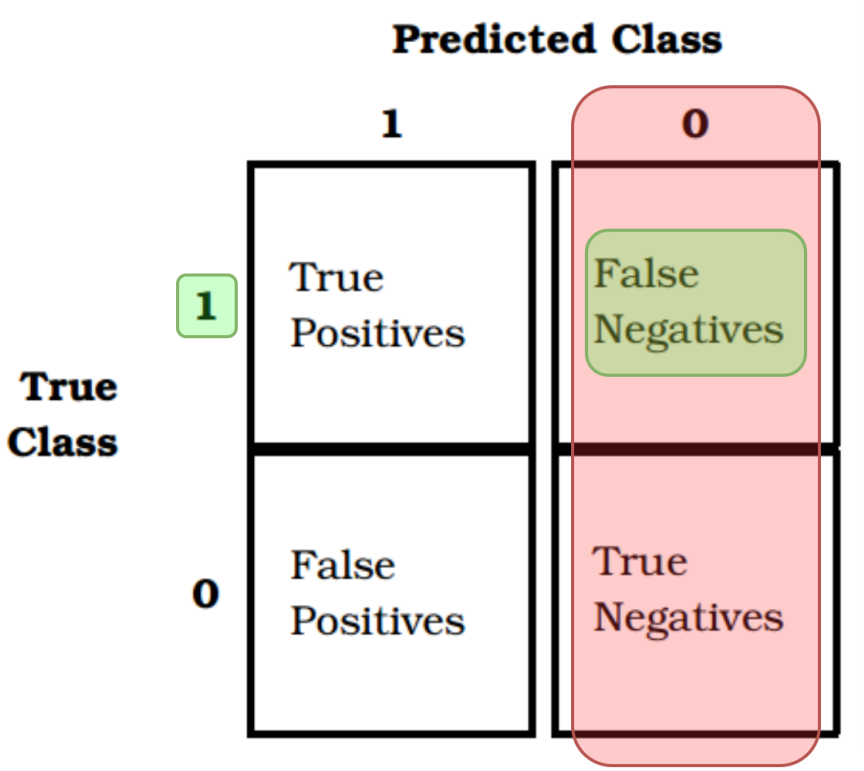
Tasa de Falso Descubrimiento - False Discovery Rate (FDR)



$$FDR = \frac{FP}{FP + TP}$$

Fracción de individuos que el modelo predice que tendrán una etiqueta positiva pero para quienes la etiqueta real es negativa.

Tasa de Falsa Omisión - False Omission Rate (FOR)

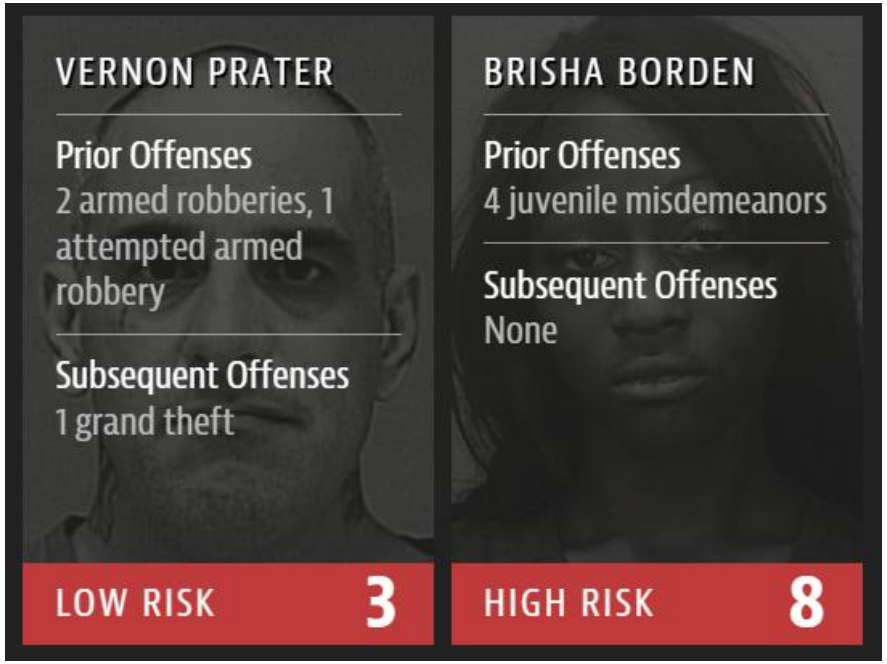


$$FOR = \frac{FN}{FN + TN}$$

Fracción de individuos que el modelo predice que tendrán una etiqueta negativa pero para quienes la etiqueta real es positiva.

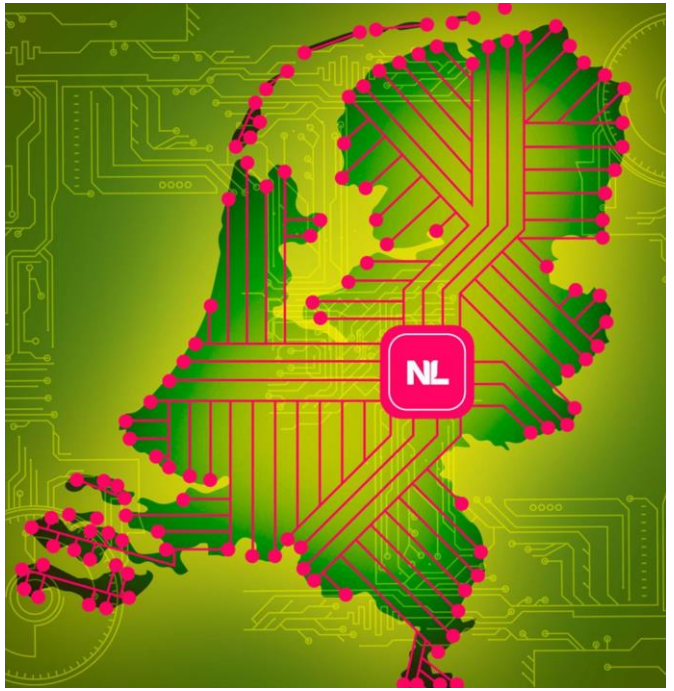
Casos de sistemas ML sesgados

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)



Machine Bias — ProPublica

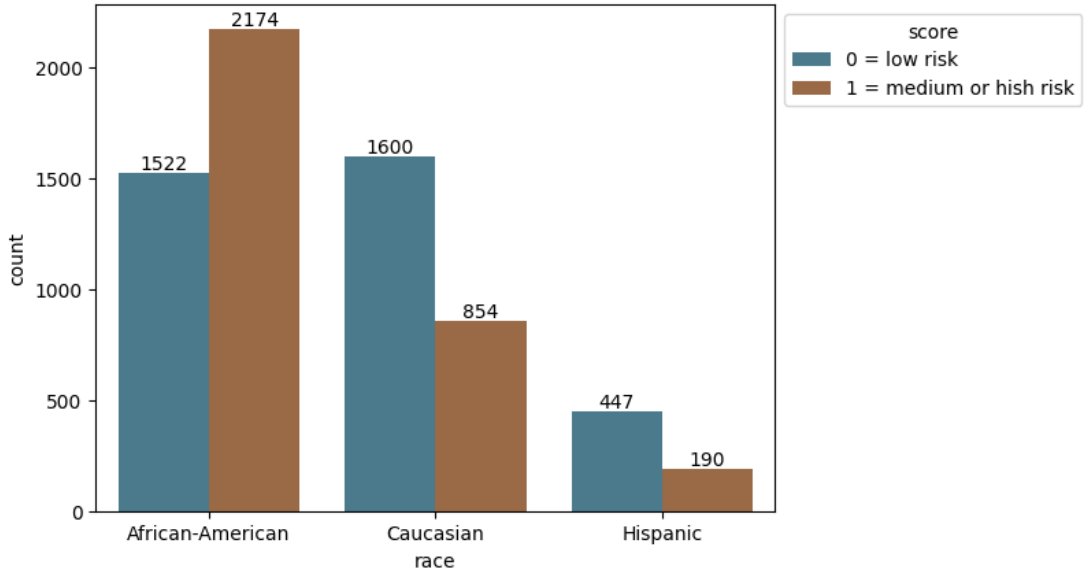
SyRI (System Risk Indication)



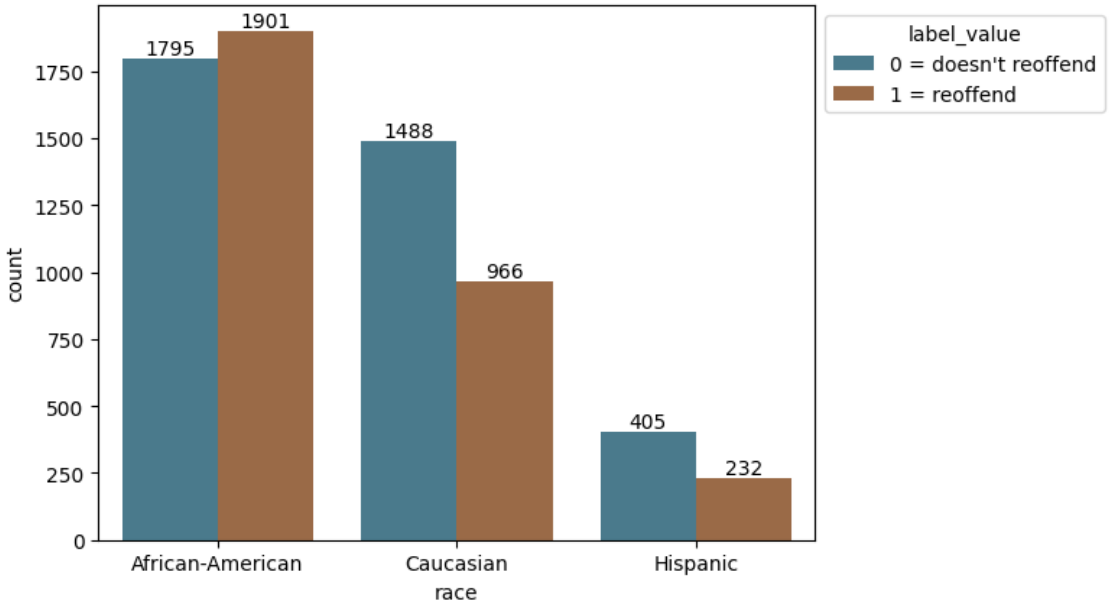
Países Bajos - Bienestar digital y derechos humanos

Sesgos en COMPAS

Predicción



Real



$$Pred_{blancos} = \frac{854}{854 + 1600} = 34.8\%$$

$$TBR_{blancos} = \frac{966}{966 + 1488} = 39.3\% \Uparrow$$

$$Pred_{negros} = \frac{2174}{2174 + 1522} = 58.8\%$$

$$TBR_{negros} = \frac{1901}{1901 + 1759} = 51.4\% \Downarrow$$

Aequitas

Conjunto de herramientas de auditoría de sesgo (open source) para auditar modelos de ML en busca de discriminación y sesgo, y para tomar decisiones informadas y equitativas sobre el desarrollo y la implementación de herramientas predictivas.

Aequitas - The Bias Report

Aequitas - Estructura de la base de datos

Debe ser un archivo CSV con la siguiente estructura

score = predicción del modelo ($\hat{Y}$)

label_value = etiqueta verdadera (Y)

attribute_n = atributos protegidos (raza, sexo, ingresos, entre otros)

score	label_value	attribute_1	...	attribute_n
1	1	A		40
1	0	B		32
0	0	A		50
0	1	B		30
1	1	B		12
	↑ Binario o continuo	↑ Binario	↑ Categorico o continuo	↑ Categorico o continuo

Selección del grupo de referencia (atributos protegidos)

Para calcular las métricas es necesario una clase de referencia, si bien no existe un estándar para su selección se recomienda lo siguiente.

1. Cuando se plantea un experimento con un grupo control la clase de referencia puede ser la que esté asociada a este.
2. En caso contrario se puede elegir como referencia la clase con más representatividad.
3. En caso de tener una base de datos balanceada la elección puede ser aleatoria.
4. El grupo de referencia puede generar diversos resultados de las métricas, por lo cual se recomienda hacer pruebas con diferentes grupos.

Aequitas - Evaluación de las Métricas

Aequitas propone un porcentaje de tolerancia que se puede modificar de acuerdo al margen de inequidad que se permitirá o tolerará en el proyecto, sin embargo, se recomienda usar el 80% para la versión WEB y para el caso de la librería de Python, un valor numérico de referencia de 1.25

Medición de sesgo en Aequitas

$$Disparity_{FNR} = \frac{FNR_{black}}{FNR_{white}}$$

Las disparidades se calculan como una proporción de una métrica para un grupo de interés en comparación con un grupo base.

Medición de equidad en Aequitas

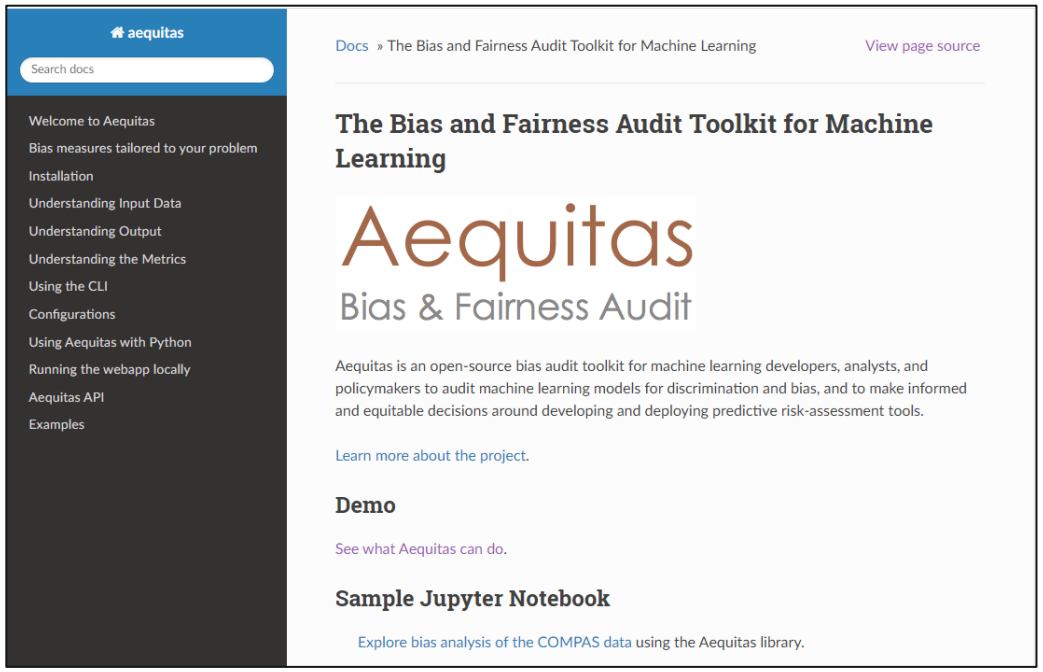
La equidad se define en relación con un grupo de referencia. En las Evaluaciones de Criterios de Equidad, un grupo cumple con la paridad si

$$(1-\tau) \leq DisparityMeasure_{Group_i} \leq \frac{1}{(1-\tau)}$$

donde τ es el umbral de equidad definido.

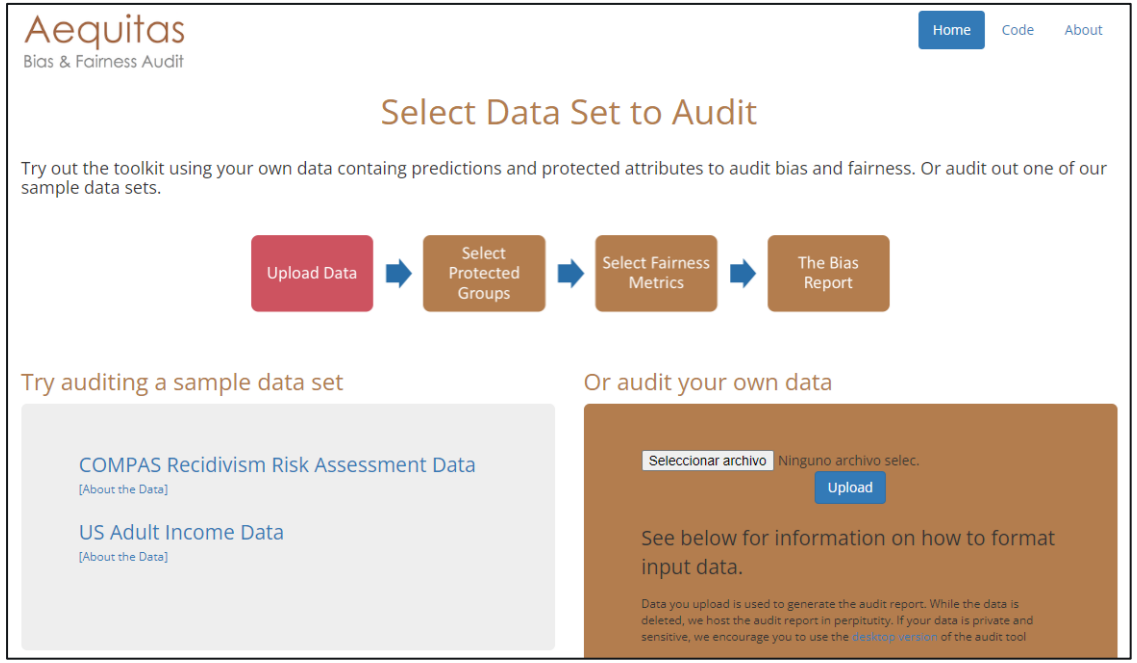
Aequitas

API de Python



[Aequitas Documentation]

Aplicación Web



[Aequitas - Webapp]

Modelos ML de carácter Punitivo

- Cuando la aplicación de un modelo de riesgo tiene un carácter punitivo, los individuos pueden resultar perjudicados al ser incluidos incorrectamente en la población de “alto riesgo” (etiqueta positiva) que recibe una intervención.

- En un caso extremo, podemos pensar en esto como detener incorrectamente a una persona inocente en la cárcel.

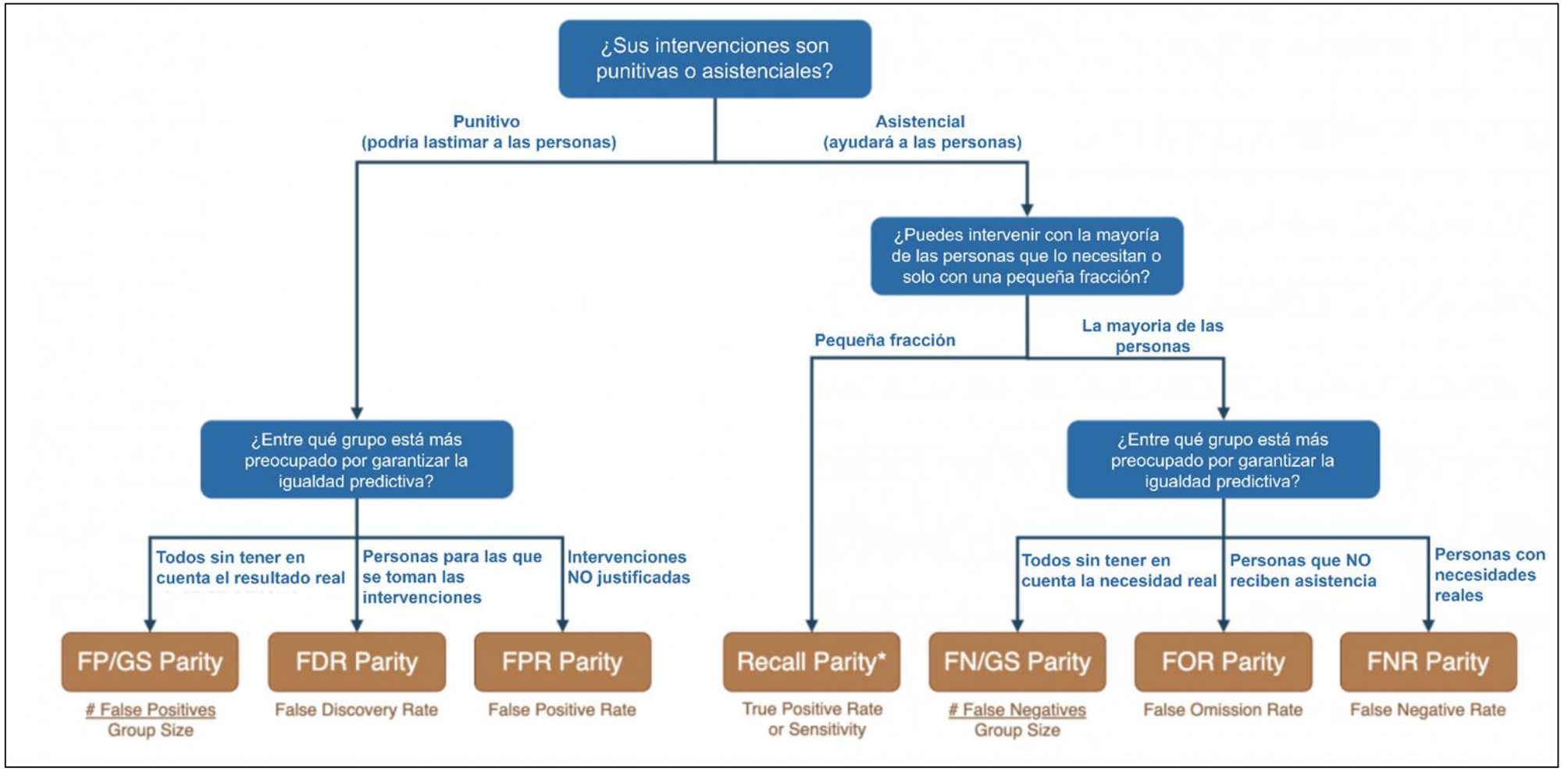
- Con las intervenciones punitivas, nos enfocamos en métricas de sesgo y equidad basadas en **falsos positivos.**

Modelos ML de carácter Asistencial

- A diferencia del caso punitivo, cuando la aplicación de un modelo de riesgo es de naturaleza asistencial, los individuos pueden verse perjudicados al ser incorrectamente excluidos de la población de “alto riesgo” que recibe una intervención.

- Mientras que el caso punitivo se centró en errores de inclusión a través de falsos positivos, la mayoría de las métricas de interés en el caso de asistencia se centran en análogos que miden errores de omisión a través de **falsos negativos**.

Aequitas - Selección de Métricas (Árbol de Equidad)

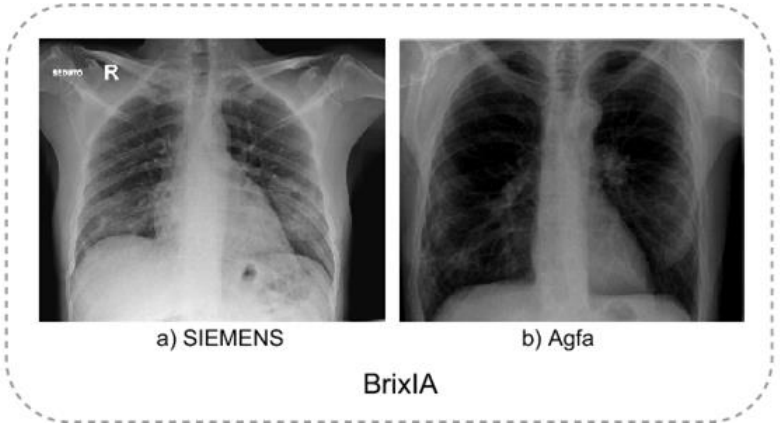


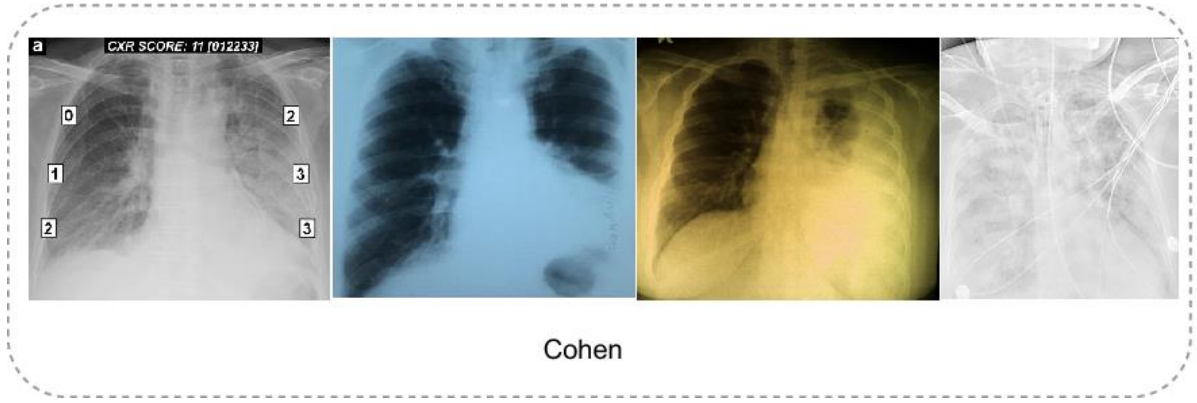
Fairness Full Tree

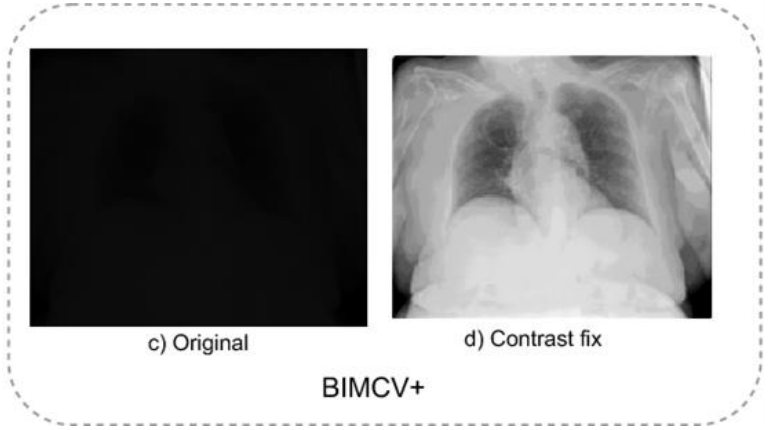
Sesgos para la detección de Covid 19

Sesgos en las BASES DE DATOS de Covid-19

- **Información del paciente:** Sexo, edad, distribución del paciente, características demográficas.
- **Condiciones de captura:** dispositivo usado, buena captura, cables o tubos.
- Desbalances y mezcla de dataset.







Sesgos en las BASES DE DATOS de Covid-19

BIMCV
Medical Imaging Databank of the Valencia Region

Covid-19 negativo
5469 imágenes

Covid-19 positivo
12802 imágenes

Métrica	Mejor Porcentaje	Modelo
Precisión	99.75%	BraMa-Net
Sensibilidad	100%	BraMa-Net
Especificidad	99.50%	BraMa-Net

Sesgo: UCI, dispositivo, temporal, segmentación

S.E.S Hospital Universitario de Caldas
Queremos devolver sonrisas
ACREDITACION EN SALUD

Covid-19 negativo
4230 imágenes

Covid-19 positivo
3077 imágenes

Métrica	Mejor Porcentaje	Modelo
Precisión	87.73%	InceptionV3
Sensibilidad	89.38%	ResNet152V2
Especificidad	92.08%	VGG16

Cov-caldas: A new COVID-19 chest X-Ray dataset from state of Caldas-Colombia | Scientific Data (nature.com)

Sesgos en los MODELOS de IA para detección de Covid-19

Preparación de la Base de datos

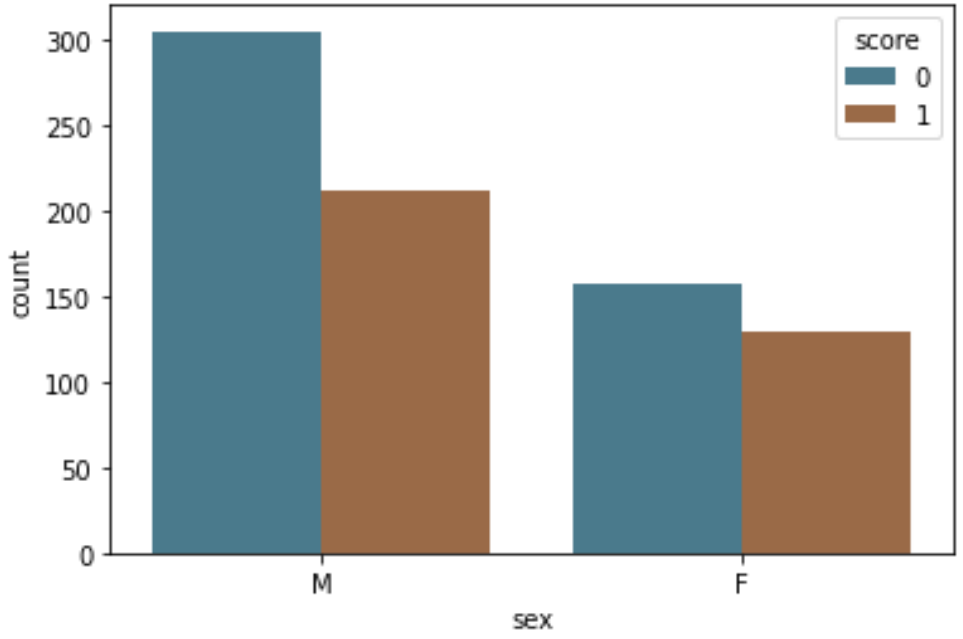
Luego de usar un modelo convolucional (VGG19) para clasificación de COVID-19 en radiografía de Tórax (Cohen dataset) se generó el siguiente CSV con las columnas requeridas y 5 atributos, el resultado de Accuracy del modelo fue del 96%

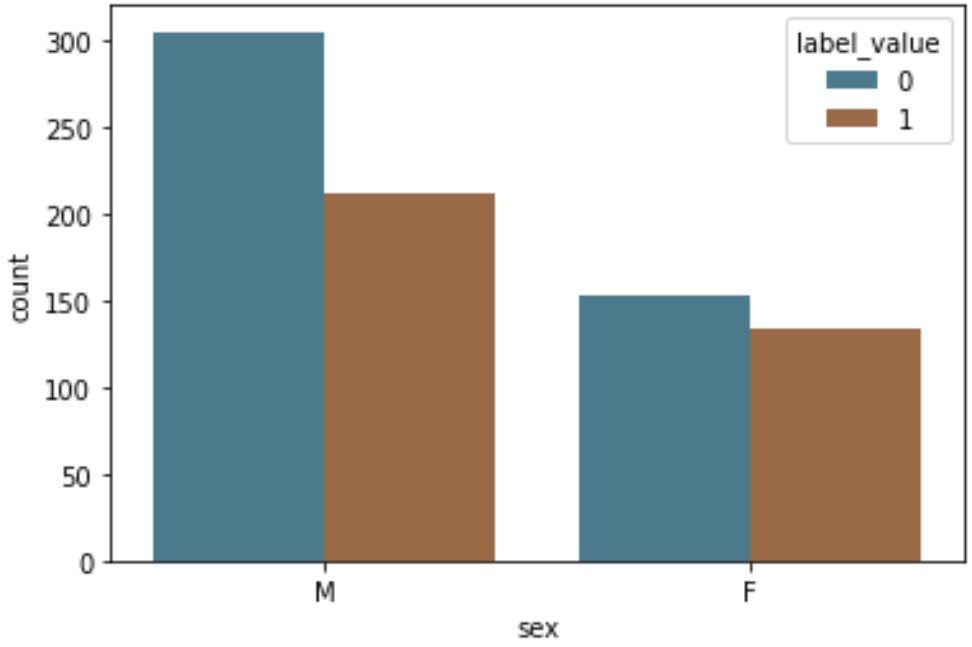
	score	label_value	age	sex	went_icu	location	date
0	0	0	65.0	M	N	Vietnam	2020
1	0	0	65.0	M	N	Vietnam	2020
2	0	0	65.0	M	N	Vietnam	2020
3	0	0	65.0	M	N	Vietnam	2020
4	0	0	52.0	F	N	Taiwan	2020

Cohen experiment.ipynb - Colaboratory

Distribución de los atributos

Luego de pre procesar la base de datos quedaron las siguientes distribuciones para el Atributo “sex”





Índice de nomenclatura para las métricas

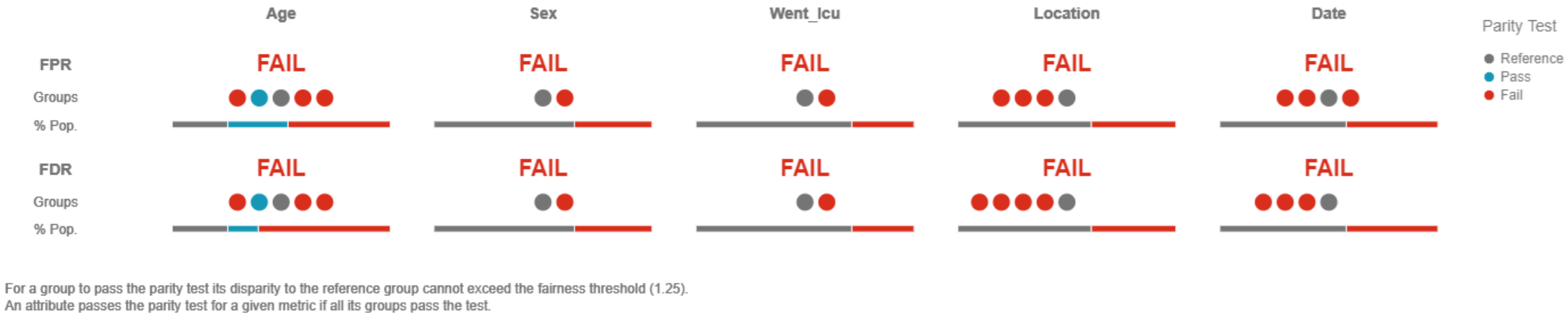
Las métricas dentro de este proyecto se ven simplificadas por la siguiente nomenclatura

Count Type	Column Name
False Positive Count	'fp'
False Negative Count	'fn'
True Negative Count	'tn'
True Positive Count	'tp'
Predicted Positive Count	'pp'
Predicted Negative Count	'pn'
Count of Negative Labels in Group	'group_label_neg'
Count of Positive Labels in Group	'group_label_pos'
Group Size	'group_size'
Total Entities	'total_entities'

Metric	Column Name
True Positive Rate	'tpr'
True Negative Rate	'tnr'
False Omission Rate	'for'
False Discovery Rate	'fdr'
False Positive Rate	'fpr'
False Negative Rate	'fnr'
Negative Predictive Value	'npv'
Precision	'precision'
Predicted Positive Ratio$_k$	'ppr'
Predicted Positive Ratio$_g$	'pprev'
Group Prevalence	'prev'

Resultados

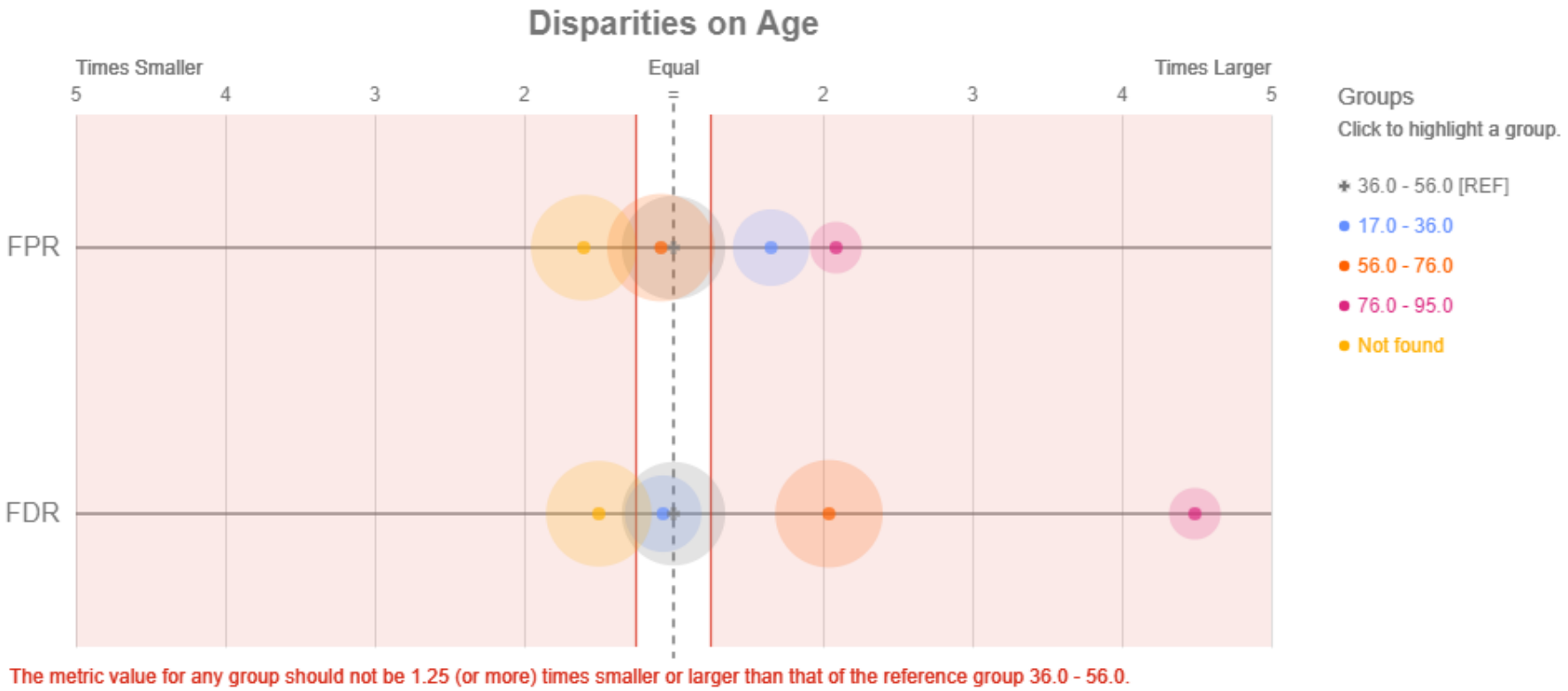
Resumen de las métricas con el dataset de Covid, este es interactivo cuando se genera por lo que se puede ver cuánto se desfasó cada clase para pasar la prueba de la métrica, con la tolerancia dada.



[Aequitas Notebook .ipynb - Colaboratory]

Disparidades para un atributo específico

```
ap.disparity(bdf, metrics, 'age', fairness_threshold = disparity_tolerance)
```



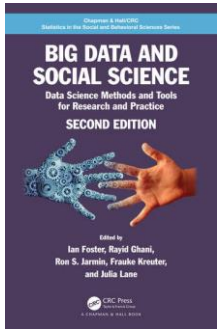
Aequitas - Ejemplos prácticos

Google colab

Enlace a COMPAS con Aequitas (Notebook)

Enlace a Detección de Covid-19 con Aequitas (Notebook)

Enlace a Artículo - Sesgos asociados datos Covid-19



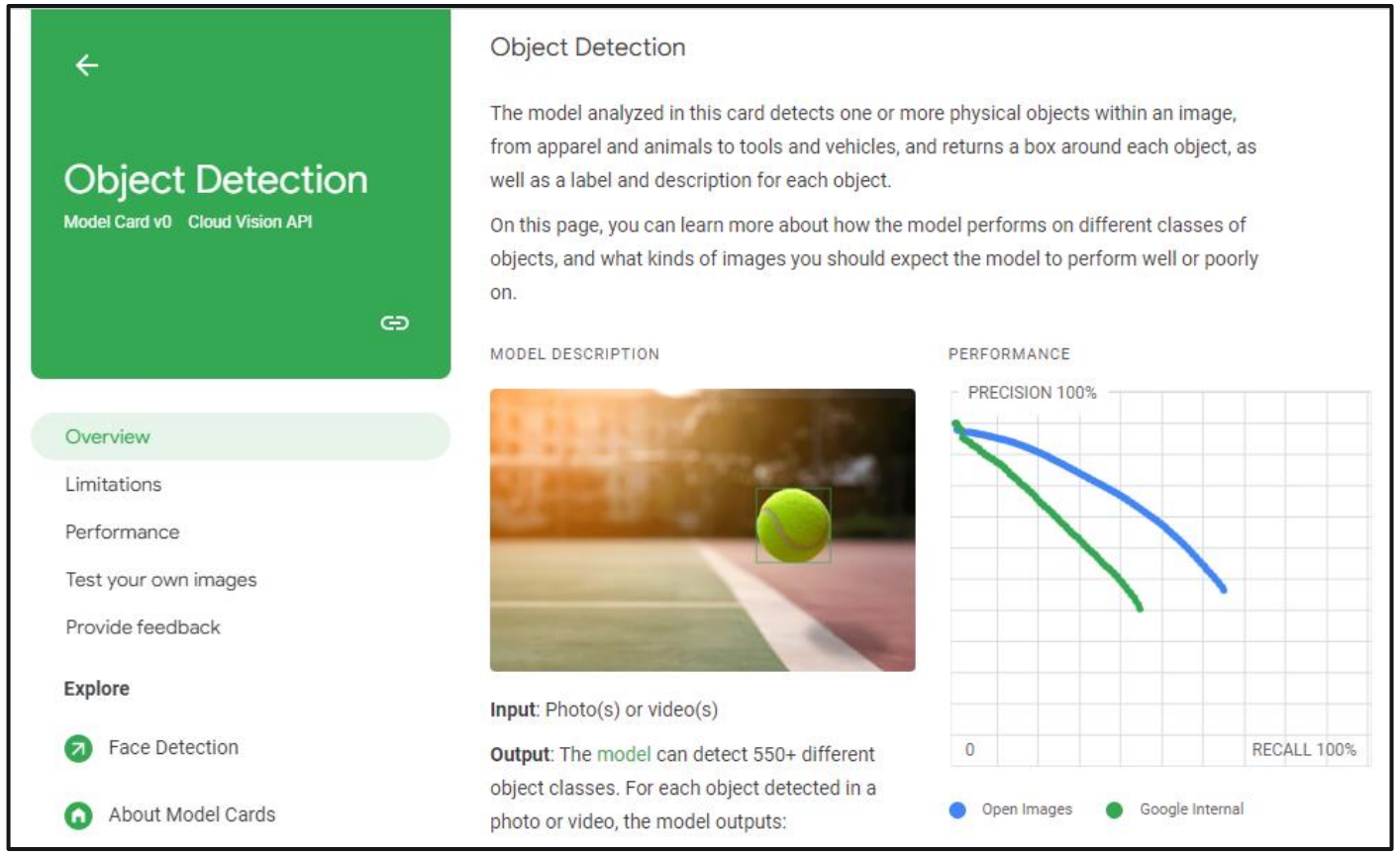
Big Data and Social Science. Capítulo 11: Sesgo y equidad

Transparencia y Análisis de Contrafactuales

Transparencia

Model Cards - Estructura

- Detalles del Modelo
- Uso Previsto
- Factores
- Métricas
- Datos de Entrenamiento y de Evaluación
- Análisis Cuantitativo
- Consideraciones Éticas
- Advertencias y Recomendaciones



[Google Cloud Model Cards - Object Detection]

SMART + SUSTAINABLE

Ejemplo de Estructura

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

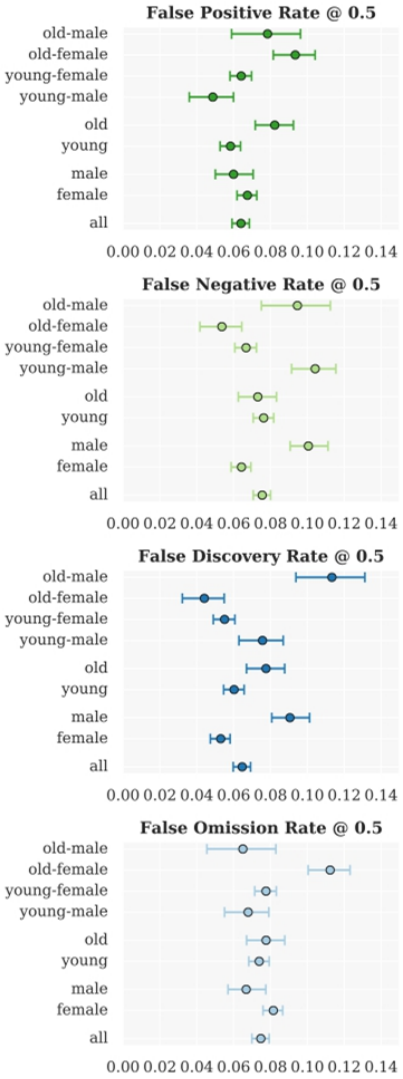
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

CULTAD DE GENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Ejemplos prácticos

Classification COVID-19 (Notebook)

Breast Cancer Dataset (Notebook)

Cats vs Dogs Dataset (Notebook)

Otro ejemplos

Model Reports - Google

Model Card - Salesforce

Model Card - Wikipedia

Models and Model Cards - MediaPipe

Model Card GPT-3 - OpenAI

Model Card InstructGPT - OpenAI

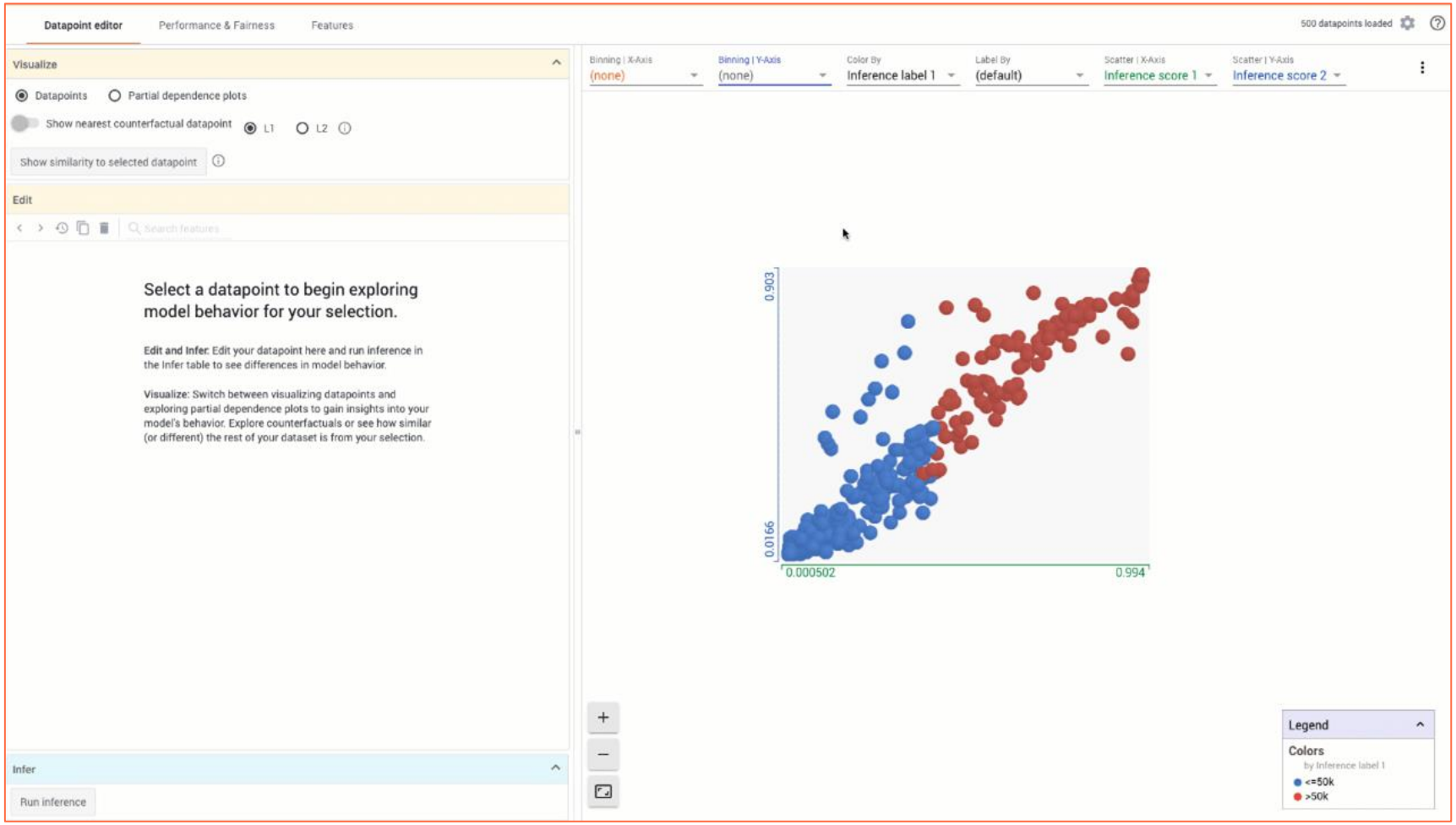
Stable Diffusion v1 Model Card

What-If Tool

Datapoint editor

Performance para modelos de regresión y multiclase o **Performance & Fairness** para modelos de clasificación binaria

Features



[Tour - Tutorial]

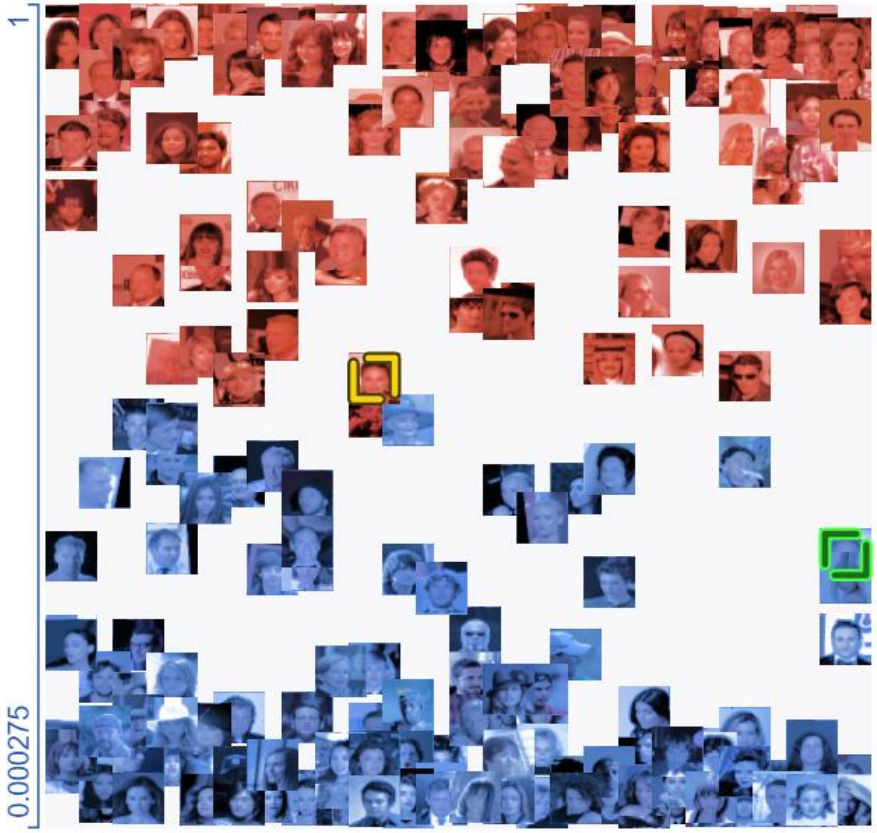
What If Tool - Ejemplo: Detector de Sonrisas

Nearest counterfactual ⓘ ○ L1 ○ L2 ◉ Custom distance

Create similarity feature ⓘ

Feature	Value(s)	Counterfactual value(s)
image/encoded		
5_o_Clock_Shadow	No 5 o'clock shadow	No 5 o'clock shadow
Arched_Eyebrows	Arched eyebrows	No arched eyebrows
Bags_Under_Eyes	No bags under eyes	No bags under eyes
Bald	Not bald	Not bald

Run	Label	Score	Delta	Run	Label	Score	Delta
1	1 (Smiling)	0.548		1	0 (Not smiling)	0.679	
1	0 (Not smiling)	0.452		1	1 (Smiling)	0.321	



Dato que genera el cambio

Ejemplos prácticos

Enlaces a What-If Tool con UCI Census Income Data Set (Notebook) (Web Demo)

Enlaces a What-If Tool - Detección de sonrisas (Notebook) (Web Demo)

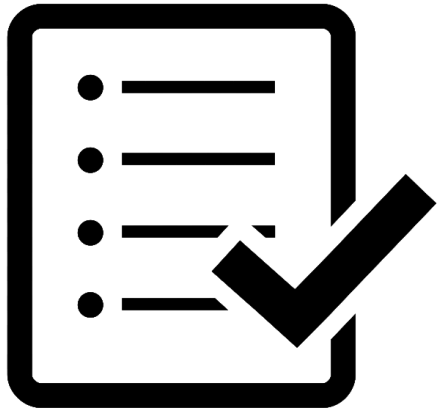
Enlace a comparación de modelos What-If Tool - Toxicity Text (Notebook)

Recursos

What-If Tool - Guías de uso

Repositorio en GitHub

Algunas aplicaciones en Chile



SMART + SUSTAINABLE

BID | LAB

FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

ALGORITMOS ÉTICOS | RESPONSABLES Y TRANSPARENTES

APOYAN:

Ministerio de Ciencia, Tecnología, Conocimiento e Innovación
Gobierno de Chile

gob digital CHILE

Dirección ChileCompra

consejo para la Transparencia

Defensoría Sin defensa no hay Justicia

Fonasa

IPS Chile Atiende
INSTITUTO DE PREVISIÓN SOCIAL

Magical

Desarrollo de Algoritmos Éticos en base a datos para la predicción de salidas judiciales favorables para la Defensoría Penal Pública

Objetivo general: Desarrollar modelos de Machine Learning para la predicción de salidas favorables en juicios, que puedan ser utilizados por la Defensoria Penal Pública, y que además sean éticamente responsables para evitar sesgos en las predicciones.

Herramientas:

Aequitas (Análisis de equidad del modelo), **What If Tools** (Análisis de contrafactuales), **Model Cards** (Transparencia), **SHAP Values** (Explicabilidad del modelo)

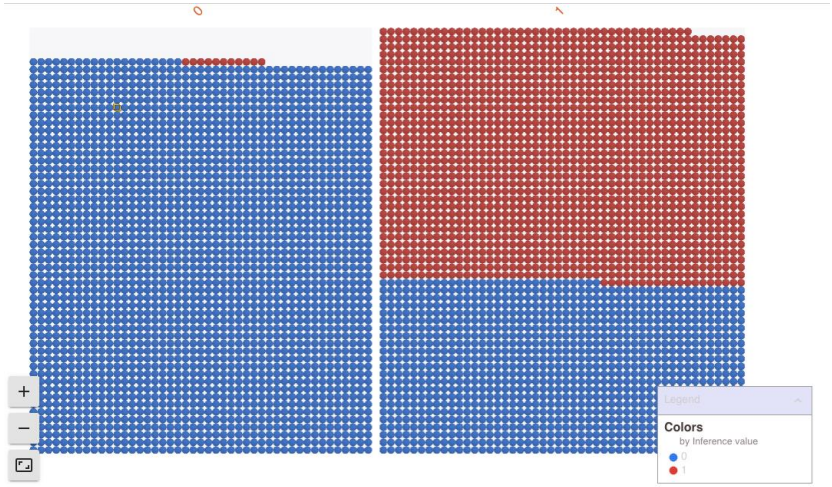
Modelo: Dataset tráfico de drogas (red neuronal).

Accuracy: 78% - **Precision**: 100% - **F1**: 74%

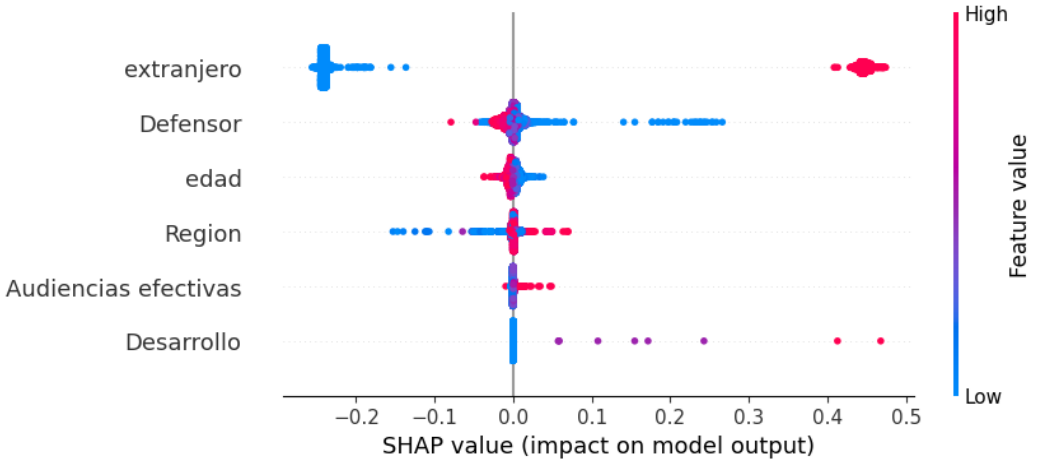
Recall: 59% Alta tasa de falsos negativos (muchas salidas no favorables, cuando debieron haber sido favorables en realidad).

- **Resultados:** Aequitas (Modelo no es equitativo)

- **Resultados:** WIT (Alta tasa de falsos negativos)



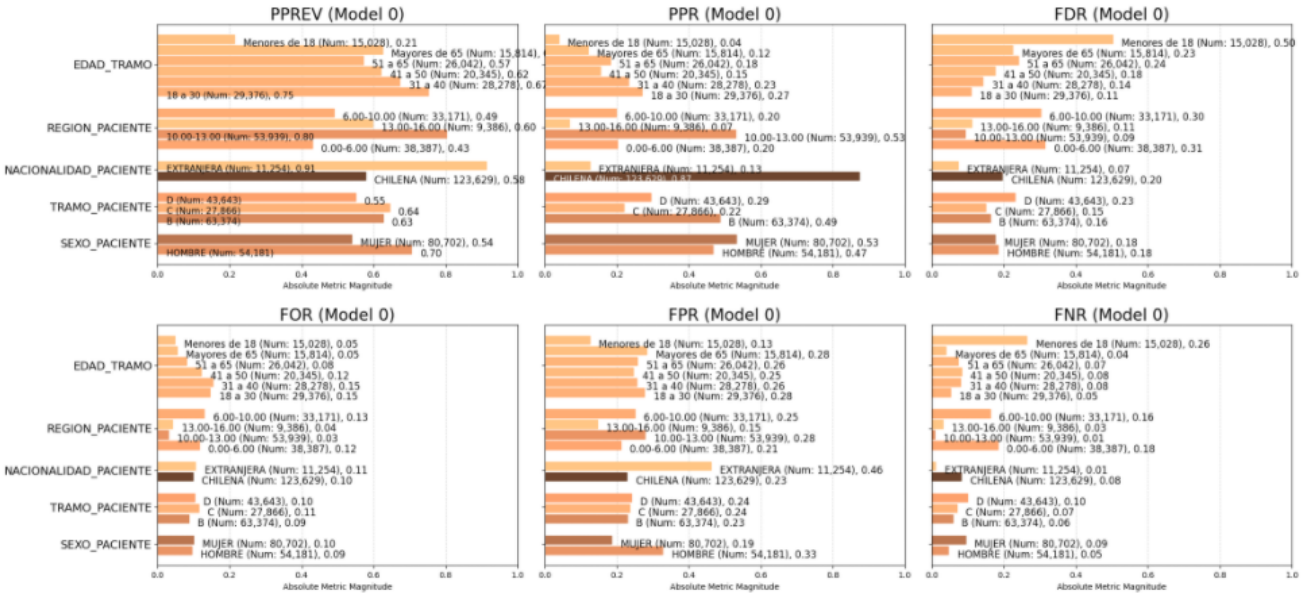
- **Resultados:** SHAP (Defensor impacta más en resultado)



- **Cómo mejorar:** Cambiar *threshold* de predicción reduce la tasa de falsos negativos. En particular, al reducirlo de 0.5 (por defecto) a 0.3, la FNR se reduce de un 40% a un 14%. Esto impacta a otras métricas, pero como nos interesa equidad en los resultados, es una importante mejora.

Fiscalización de Prestadores de Salud con ayuda de ML - FONASA

- **Objetivo general:** Diseñar y desarrollar una solución de inteligencia artificial y/o ciencia de datos que identifique a los prestadores que presentan comportamiento sospechoso de fraude y seleccione la muestra de prestaciones de salud a fiscalizar.
- **Herramientas**:
  - **Aequitas**: Análisis de equidad del modelo
- **Modelos:**
  - Random Forest, Gradient Boosting Machine, Extra Trees
- **Resultados:** Aequitas (Modelos no son equitativos) y presentan sesgos en atributos como sexo y nacionalidad del paciente, asi como el monto de facturación.




```
In [72]: gof = f.get_overall_fairness(fdf)
         gof

Out[72]: {'Unsupervised Fairness': False,
          'Supervised Fairness': False,
          'Overall Fairness': False}
```

IPS
Ministerio del Trabajo y Previsión Social

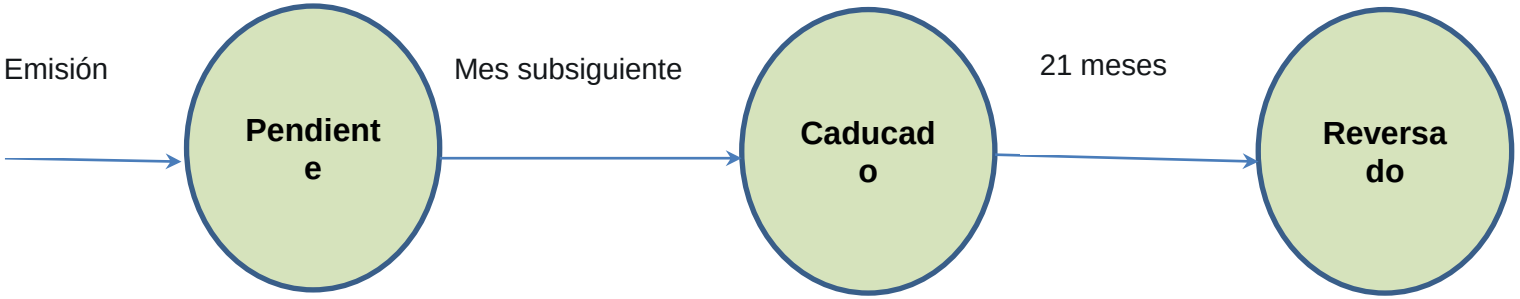
Gobierno de Chile

PROYECTO
NO COBROS

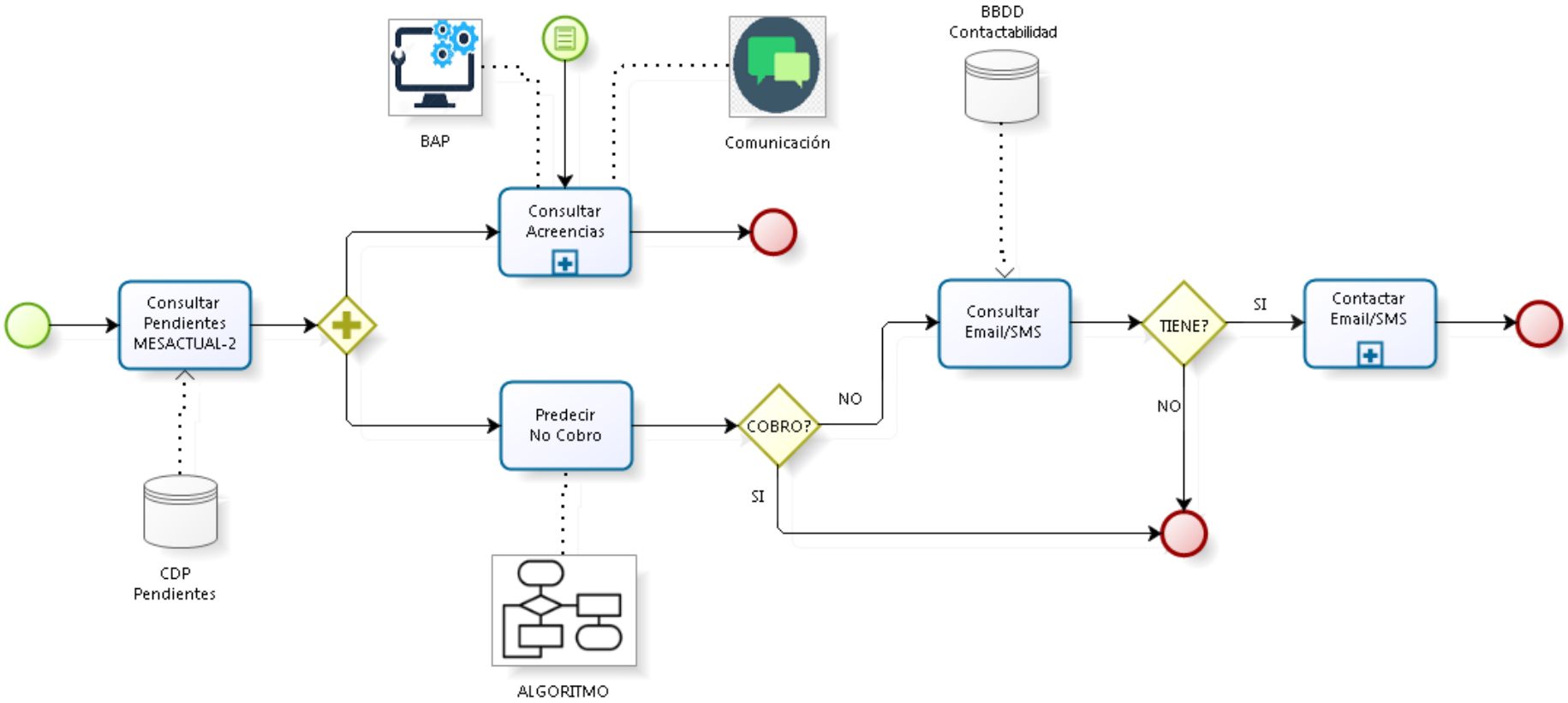
Plan piloto
Modelo Predictivo de No Cobro PGU

Contexto

- Los No Cobros son documentos de pago emitidos que después de 3 meses sin ser cobrados pasan a estado «caducado», y al cumplir 24 meses deberían ser «reversados».
- Actualmente existen aproximadamente 800 mil documentos caducados.
- A pesar que no existe imperativo legal de gestionar los no cobros con los beneficiarios, si existe un imperativo ético, por ello proponemos desarrollar una segunda etapa durante 2023, centrada en la implementación de las soluciones diseñadas.



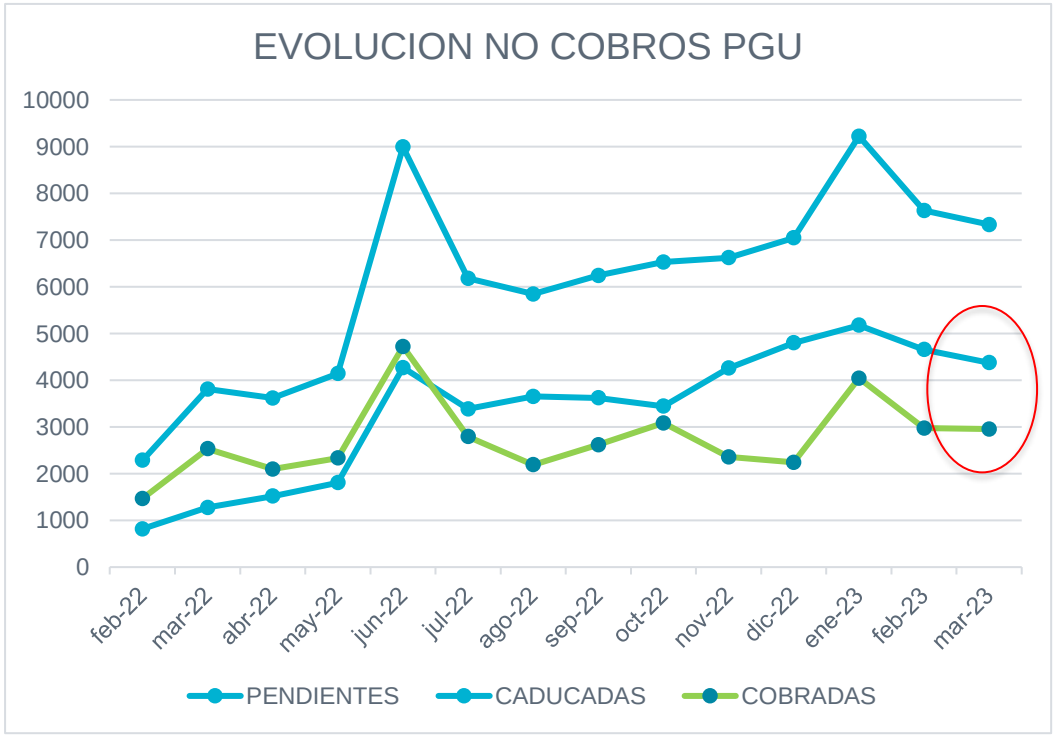
PROCESO GESTION DE PAGOS NO COBRADOS A PUNTO DE CADUCAR



PLAN PILOTO MODELO PREDICTIVO NO COBRO PGU

- Objetivo: disminuir los PGU caducados.
- Los datos de entrenamiento: 70.564 PGU pendientes hasta el tercer mes, entre febrero de 2022 y enero de 2023.
- Los datos para predecir: 7.334 PGU no cobradas a punto de caducar de marzo de 2023.

MES	PENDIENTES	CADUCADAS	COBRADAS	%NOCOBRO
feb-22	2289	819	1470	35,8
mar-22	3814	1279	2535	33,5
abr-22	3622	1522	2100	42,0
may-22	4144	1809	2335	43,7
jun-22	8993	4270	4723	47,5
jul-22	6183	3387	2796	54,8
ago-22	5847	3655	2192	62,5
sep-22	6244	3623	2621	58,0
oct-22	6533	3446	3087	52,7
nov-22	6623	4261	2362	64,3
dic-22	7048	4803	2245	68,1
ene-23	9224	5178	4046	56,1
feb-23	7635	4659	2976	61,0
mar-23	7334	4379	2955	59,7



MATRIZ DE CONFUSION DE MODELOS

- La clase positiva es el NO COBRO

MATRIZ DE CONFUSION	TP	FN	TN	FP
REGRESION LOGISTICA	5.900	1.748	3.793	2.672
RANDOM FOREST	5.596	2.052	4.319	2.146
RED NEURONAL MLP	5.462	2.186	4.346	2.119
SUPPORT VECTOR MACHINE LINEAL	4.344	3.304	4.975	1.490
SUPPORT VECTOR MACHINE RADIAL	5.642	2.006	4.157	2.308

ACCURACY Y SENSITIVITY DE MODELOS

RATES MODELOS	ACCURACY	SENSITIVITY
REGRESION LOGISTICA	68,8%	77,1%
RANDOM FOREST	70,3%	73,2%
RED NEURONAL MLP	69,5%	71,4%
SUPPORT VECTOR MACHINE LINEAL	66,0%	56,8%
SUPPORT VECTOR MACHINE RADIAL	69,4%	73,7%

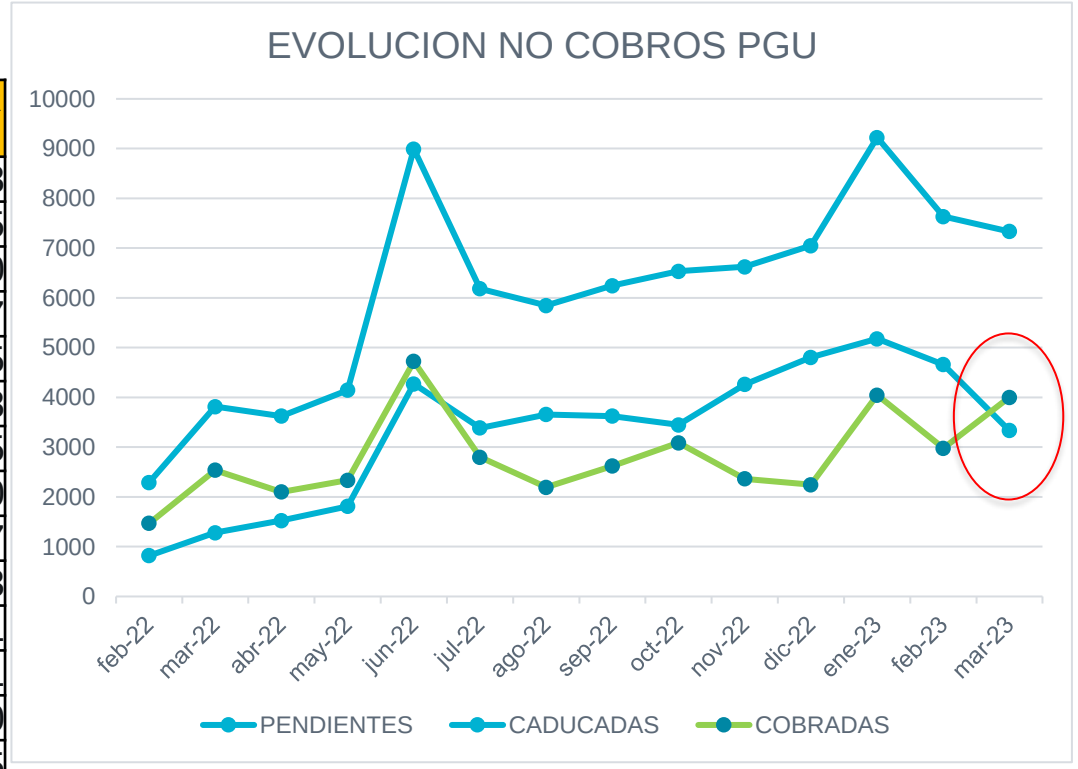
		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

- Aprendimos que para este caso de negocio el indicador relevante es el **Sensitivity**.

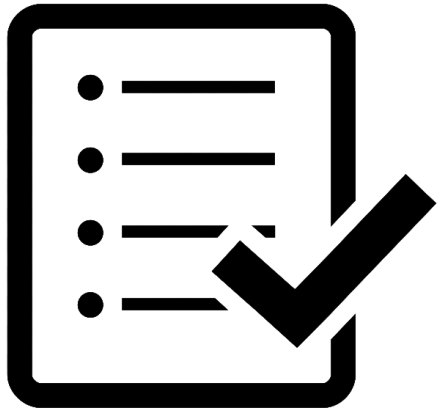
RESULTADOS OBSERVADOS

- El plan piloto se efectuó para pagos PGU pendientes emitidos en marzo de 2023 y que en mayo aun no eran cobrados, y este último mes la curva de COBRADAS cruza la de CADUCADAS debido a la acción del Modelo Predictivo

MES	PENDIENTES	CADUCADAS	COBRADAS	%NOCOBRO
feb-22	2289	819	1470	35,8
mar-22	3814	1279	2535	33,5
abr-22	3622	1522	2100	42,0
may-22	4144	1809	2335	43,7
jun-22	8993	4270	4723	47,5
jul-22	6183	3387	2796	54,8
ago-22	5847	3655	2192	62,5
sep-22	6244	3623	2621	58,0
oct-22	6533	3446	3087	52,7
nov-22	6623	4261	2362	64,3
dic-22	7048	4803	2245	68,1
ene-23	9224	5178	4046	56,1
feb-23	7635	4659	2976	61,0
mar-23	7334	3334	4000	45,5



Conclusiones y Recomendaciones



- **Medición de Impacto**
  - **Algorithm Impact Assessment:** Identifica el nivel de impacto del proyecto dependiendo del área y mitigaciones. *Framework*

- **Identificación de sesgos**
  - **Aequitas:** Audita los sesgos que un modelo puede tener identificando diferentes tipos de equidad. *Librería*
  - **AI Fairness 360 - Fairlearn:** Audita sesgos en modelos y datasets, adicionalmente mitiga los sesgos encontrados. *Librerías*

- **Transparencia y Explicabilidad**
  - **Model Cards:** Informa a los usuarios sobre lo que un modelo puede y no puede hacer, así como los tipos de errores que cometen. *Framework*
  - **What-If Tool:** Examina el rendimiento en situaciones hipotéticas, analiza la importancia de diferentes características de datos y visualiza el comportamiento del modelo. *Software*

Retos de aspectos éticos en IA

- Desconocimiento de los conceptos y herramientas.
- Complejidad de uso de las herramientas.
- Limitaciones de las herramientas.
- Falta de difusión.
- Falta de regulación.

SMART + SUSTAINABLE

FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Gonzalo Ruz, Ph.D

Mariana German,
Estudiante Msc

Joshua Bernal,
Ing Biomédico

Gracias!!!
r.tabares@uai.cl

Ricardo Ortega,
Ing Biomédico

Nelson Salazar,
Estudiante Msc